Optimized data processing algorithms for biomarker discovery by LC-MS

Christin

Optimized Data Processing Algorithms for Biomarker Discovery by LC-MS

Christin

Paranymphs: Tejas P. Gandhi
Tita A. Listrowardojo

The work described in this thesis was performed in the research group Analytical Biochemistry, Faculty of Mathematics and Natural Science, University of Groningen and within the graduate school GUIDE. The project was funded by the Netherlands Bioinformatics Center (BioRange 2.2.3).

Cover art created using www.juliasets.dk/Julia.htm
Printed by: Ridderprint Offsetdrukkerij B.V.

RIJKSUNIVERSITEIT GRONINGEN

Optimized Data Processing Algorithms for Biomarker Discovery by LC-MS

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
woensdag 11 mei 2011
om 11.00 uur

door

Christin

geboren op 15 augustus 1981
te Tangerang, Indonesië

Promotores:	Prof.dr. R.P.H. Bischoff
	Prof.dr. A.G.J. van der Zee
	Prof.dr. A.K. Smilde
Copromotores:	Dr. P.L. Horvatovich
	Dr. H.C.J. Hoefsloot
Beoordelingscommissie:	Prof.dr. E. T. Hankemeier
	Prof.dr. O. Kohlbacher
	Prof.dr. E.R. van den Heuvel

ISBN: 978-90-367-4914-5.
ISBN (digital version): 978-90-367-4915-2

Table of Contents

Chapter 1

General Introduction

Christin C, Bischoff R, Horvatovich P. *Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery.* Talanta. 2011 Jan 30;83(4):1209-24.

The recent widespread application of mass spectrometry to quantify and identify large numbers of compounds in biological matrices leads to an explosion of acquired data. The goal of these measurements is to explore the underlying molecular mechanism of disease, to identify compounds (biomarkers) strongly related to the stage of the disease, its onset or progression for diagnostic purposes, to identify novel drug targets, and to follow the efficiency of treatment. The dynamic behavior of multifactorial diseases requires a systems biology approach to find reliable biomarkers taking molecular regulatory mechanisms, compound flux and concentration changes into account [1]. To explore robust changes in molecular systems related to disease, it is necessary to analyse a large number of samples from different biological entities, for example from different, clinically well-characterised patient groups. Generally biomarker research is based on complex biological samples containing a large number of diverse compounds such as proteins, peptides and metabolites.

Liquid chromatography coupled to mass spectrometry (LC-MS) is one of the most widely used comprehensive profiling techniques to measure compounds in biological materials. A single comprehensive LC-MS analysis cannot cover all types of compounds in the samples. Instead, it measures one class of compounds such as metabolites, lipids and proteins, leading to biomarker discovery in this class of molecules. Even with a technique targeting one of the afore-mentioned classes of compounds, not all types of molecules can be measured due to ionization limitations of the electrospray interface. Another challenging problem is the wide dynamic concentration range of the compounds, which can reach 9-11 orders of magnitude in the case of body fluids such as blood [2, 3]. From this wide dynamic concentration range, modern mass spectrometers are only able to cover 2-4 orders of magnitude. The gap between the existing and measurable dynamic concentration range can be either reduced by using comprehensive fractionation (4-6 orders of magnitude), multidimensional chromatography (up to 8 orders of magnitude) [4] or by targeting a specific subclass of compounds, *e.g.* by using an affinity enrichment step of a certain type of glycoproteins on a lectin column (up to 5-7 orders of magnitude) [5]. One other challenging factor is that, although proteins and protein complexes are directly involved in the molecular processes of biological phenomena, it is their peptide constituents obtained after enzymatic cleavage that are actually measured since they are more suitable for liquid chromatography analysis and they have better ionization properties than intact proteins or protein complexes. The most widely used endopeptidases cut proteins at well-defined sequence positions, resulting in non-overlapping peptides mixtures, from which only a fraction of theoretical possible peptides are detected. In this peptide-centric approach also called as "bottom-up", or "shotgun" strategy, the quantity of initial proteins are therefore determined indirectly based on few or more peptides, which lead to misleading quantification and identification in case of the presence of multiple highly homologous proteins having one or few peptides in common, proteins with multiple splice variants, proteins presenting different degrees of post-translation modifications (PTM) or in case of the presence of various truncated forms of the same protein [6, 7].

Biomarker discovery requires close collaboration between medical researchers, analytical chemists and bioinformaticians in order to obtain the relevant molecular information related to different aspects of diseases [8, 9]. This includes patient cohort selection, sampling of biological material, sample storage, sample preparation,

choice and optimization of LC-MS profiling platform, data analysis providing protein identifications, quantification, statistical analysis and experimental validation of the results. Several review articles describe the various techniques and steps of protein profiling for biomarker discovery in detail [9, 10].

Bioinformatics plays an important role in this process as it has the goal to extract quantitative and qualitative information for a large number of compounds (proteins and metabolites) that are present in complex biological samples and to select the discriminatory compounds between predefined sample sets. Recent advances in sample preparation methods, in liquid chromatography and in mass spectrometry instrumentation resulted in a large diversity of acquired data. This results in a huge challenge for bioinformatics to provide reliable information extraction and knowledge generation approaches. The computational tools must evolve continuously to keep up with the different types of generated data. Besides direct information extraction and knowledge discovery from raw data, bioinformatics plays an important role in experimental design, quality assessment of the profiling platform, sampling methods, sample handling, storage and preparation methods, or quality control of data pre-processing, statistical analysis and statistical validation.

This chapter focuses on fundamental data processing and current challenges in supporting biomarker discovery research in proteomics for diagnosis and treatment follow-up using LC-MS of label-free, shotgun proteomics data, highlighting significant innovations in the bioinformatics field such as new algorithms, data integration, high throughput automatic data preprocessing solutions, quality control of different data processing modules and complete workflows, including assessment of the quality of sample preparation steps and LC-MS profiling platforms [9, 11-19]. We will also investigate how insights from analytical chemistry contribute to parameter optimization leading to the development of novel bioinformatics applications that provide more accurate and reliable information extraction from the raw data. Alternative approaches based on differential labeling of samples with reagents having the same chemical but different stable isotope constitution have been covered in other reviews [20-27] and will not be treated here. This chapter limits the discussion further to biomarker discovery aiming to determine comprehensively the identity and the quantity of sample constituting proteins using analytical methods with low sample throughput. Biomarker validation using analytical methods with high sample throughput providing quantitative information on preselected list of proteins by using multiple reaction monitoring, antibody arrays and ELISA will not be covered here. Recommendations on analytical, clinical and informatics aspects of biomarker discovery and validation as well as their limitations were discussed recently in several reviews [28-34].

1. DATA PROCESSING PIPELINES IN LC-MS

LC-MS has become the major platform for analyzing samples in biomarker discovery research due to its relatively high throughput (60-90 min for analysis of one sample), sensitivity, selectivity and the coverage of many peptides and proteins [9, 35, 36]. In label-free LC-MS experiments, proteins or produced tryptic peptides are not modified chemically and their isotope constitution is unchanged. Large numbers of samples are analyzed independently by LC-MS resulting in corresponding raw data files. The quantitative and compound identity information is extracted using dedicated data

processing pipelines. This is followed by matching compound quantity and identity across several chromatograms, resulting in a matrix containing quantitative information about a large number of compounds in the different samples. In shotgun proteomics approach the target compounds are proteins, therefore methods are required to determine the original protein composition of samples and their quantities based on incomplete set of measured constituting peptides. Compounds discriminating between predefined classes of samples are obtained from this matrix using dedicated statistical analysis and validation pipelines. When a system biology approach is involved in the biomarker discovery process, it is necessary to couple the list of discriminating proteins to protein interaction (e.g. STRING, BIND) or pathway (e.g. KEGG) databases [21, 37] to elucidate the disease mechanism. Figure 1 shows the main parts of a generic proteomics pipeline for biomarker discovery.

Most of the measured signals by LC-MS are not related to real compounds but are part of white noise, background ions or simply chemical noise. Different mass analyzers generate data of different structure due to differences in scanning speed, mass resolution, measured dynamic concentration range, changes in peak width and resolution across the m/z domain and varying mass accuracy [38]. The most common mass analyzers applied in proteomics biomarker research are quadrupole, 3-dimensional quadrupole ion trap, 2-dimensional linear ion trap, time of flight, and Inductively-Coupled Resonance (ICR) trap family of mass spectrometers such as Orbitrap and Fourier Transform Ion Cyclotron Resonance Mass Spectrometers (FTMS) [39]. Besides mass spectrometers may dispose different numbers of mass analysers, and could use different ionisation method such as electrospray, ionspray, matrix-assisted laser desorption ionization (MALDI) to name the most frequently used methods to analyze proteomics samples.

In label-free LC-MS proteomics experiments, there are two types of widely used mass spectrometry data. The first data contain mass spectra obtained with one mass analyzer and is referred as single-stage mass spectrometry data (MS-1) in the literature. The second type of data is heterogeneous and contains cyclic series of MS-1 and precursor ion fragmented spectra (MS/MS). Each cycle begin with MS-1 spectra, then it is followed by defined number (generally 1-10) of MS/MS spectra obtained from the most abundant ions of the MS-1 spectra. This acquisition mode is referred as data-dependent acquisition (DDA) and abbreviated as DDA MS/MS data. The reader is referred to dedicated books [38, 40, 41] and reviews [39, 42, 43] for further reading on the main characteristics of different types of mass analyzers, ionization methods and acquisition modes. Label-free quantification is a semi-quantitative method and provides information on relative quantity changes of the same compounds in different samples. For most applications such as biomarker discovery, detection of relative protein changes is sufficient information, but in system biology type of studies, the use of stable isotope labeled standard is necessary to provide absolute quantities of proteins in samples [44].

Quantitative information can be obtained from both MS-1 and DDA MS/MS data. Quantitative methods using DDA MS/MS are based on spectral counting, and use the number of MS/MS spectra that are acquired per peptide ion(s) for the quantification of a given protein. Abundant proteins generate abundant peptide fragments that have a higher probability to be selected as precursor ions for DDA MS/MS analysis. Nevertheless, despite spectral counts shows good linearity with analysed protein

amounts [45, 46], the number of MS/MS spectra per protein suffer from saturation effect, undersampling, and from the limited linear concentration range - compared to MS-1 quantification methods [47].

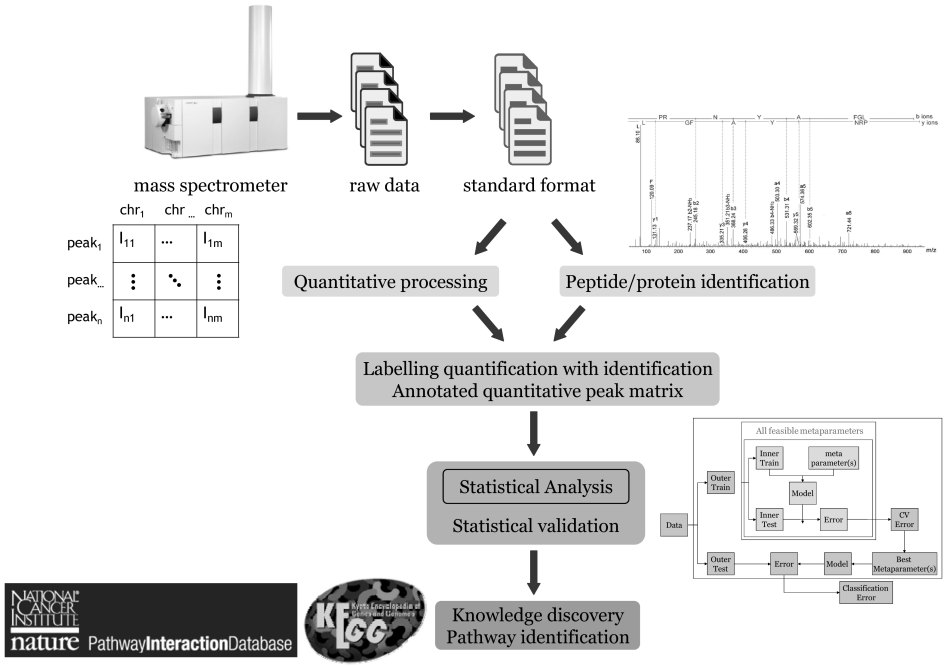

Figure 1. Main modules of a generic biomarker data processing workflow. Raw data from the mass spectrometer are converted to one of the standard data formats such as mzXML, mzData or mzML. Quantitative information and identification of proteins and peptides are performed separately from the same file or from a different data file. This is followed by labeling of the quantitative information with identifications. The statistical analysis and validation is performed on the labeled quantitative data and provides a list of discriminatory proteins that can be used for knowledge discovery with pathway analysis tools using for example KEGG (http://www.genome.jp/kegg/) or the Pathway Interaction Database (http://pid.nci.nih.gov/).

Spectral counting methods enable both absolute and relative quantification of proteins. Several bioinformatics methods use the spectral counting approach [46, 48-50]. Exponentially Modified Protein Abundance Index (emPAI) [51, 52] uses the number of identified peptides to calculate the relative molar or weight fraction of a given protein in the respective sample. Absolute Protein Expression (APEX) [53, 54] uses the measured and predicted peptide counts for quantification of peptides and proteins by considering the influence of the recovery of peptides from the cation-exchange and reversed-phase LC dimensions as well as the predicted ionization efficiency of the peptide in the ion source of a particular mass spectrometer. Recently a new method, which combines the quantification of MS-1 and MS/MS spectra by taking into account the ion count in MS-1 of the three most abundant peptides provides better quantification for proteins than spectral counting and gives the absolute protein quantity by using a single protein standard [55]. DDA MS/MS measurement is subjected to large variability regarding the

identified peptides and proteins [56], therefore more precise quantification providing a larger dynamic concentration range than spectral counting can be obtained using peptide ion counts in MS-1 data. Recently, a modified version of the MS/MS acquisition strategy called directed MS was introduced with modern high resolution Q-TOF and Orbitrap instruments. Directed MS differ from DDA MS/MS by using a different strategy to select precursor ions for fragmentation. Instead of using the most abundant signal intensity for the precursor ion selection, it performs first an MS-1 analysis and obtains an inclusion list of precursor ions with retention time window after data processing. The second MS/MS analysis is performed on precursors, which are present in the inclusion list obtained previously. This method prevents multiple reanalysis of the same peptide, allows identification of low abundant components and peptides with interesting features such as distinctive isotopic pattern, mass defect or differently modified peptides [44, 57].

Multiple Reaction Monitoring (MRM) is gaining popularity in targeted quantitative analysis for small proteomes and has the advantage to cover a large dynamic concentration range across 5 orders of magnitude [58-60]. MRM has relatively high sample throughput (30-60 min for analysis for one sample), is able to measure few hundreds of proteins in one experiment and requires to monitor 5 peptides per protein selected with the help of PeptideAtlas [61] or with prediction using bioinformatics tools such as PeptideSieve [62]. Monitoring of each proteolytic peptide requires at least 3 optimized MRM transitions selected with use of a spectral library [63, 64]. However experimental validation of the MRM transitions and their selectivity for a given problem is required to conduct reliable analysis, which can be performed by synthesis and analysis of synthetic peptide standards. Synthetic, stable isotope labeled peptide standards may be used for absolute quantification. Due to their wide dynamic concentration range, MRM-based methods can be successfully applied for validation of multiple biomarker candidates [65, 66]. A recent perspective paper describes and compares the DDA MS/MS, directed MS and MRM based proteomics analysis strategies facilitating the methodological choice for experimental researcher [44].

The large variety of raw data formats from different mass spectrometer vendors was recently standardized using several alternatives of Extensible Markup Language (XML) formats. Widely used formats are mzXML [67] (developed at the Institute for Systems Biology in Seattle Proteome Center) and mzData [68] (developed by the Human Proteome Organization Proteomics Standard Initiative or HUPO PSI). These two formats were lately merged by the HUPO-PSI [69, 70] into a new standard called mzML [71]. Several standardization attempts were made recently, mainly by HUPO-PSI, to standardize other types of proteomics data formats such as peak list, however these formats are less widely used by the proteomics community [72-74]. It is necessary for a data processing algorithm to accommodate these standardized raw data formats as the initial input of their pipeline. The standardization is not only necessary for raw data storage, but also to store intermediate results in order to enable data exchange and reusability of the results from each step in different data processing pipelines.

Label-free LC-MS data pre-processing pipelines convert the raw data into a matrix containing quantitative information on the characterized and preferably identified compounds in each of the samples amenable for statistical analysis. The main modules of such pipelines with the data flow during this conversion is presented in

Figure 2. This procedure begins with raw data pre-filtering (such as noise reduction, data reduction etc.), and is followed by detection and quantification of peaks, and results in peak lists characterized among other things by quantity, retention time and m/z value. These peak lists can be further reduced by deisotoping and summing up the intensity of compound-derived ions with different charge states. However, these steps can be also performed after the peaks have been matched across multiple chromatograms. Peptide-related peaks in different chromatograms have to be aligned or corrected in all three dimensions of MS-1 data: time alignment in the retention time dimension, mass calibration in the m/z dimension and normalization in the intensity dimension. The final step is peak matching, which has the goal to find the same peaks in multiple chromatograms and to provide the quantitative peak matrix characterized by m/z and retention time values.

Data processing pipeline should be flexible enough to adapt to the characteristics of the data sets that are dependent on pre-analytical factors, the type of mass spectrometer and experimental design of the sample preparation and sample-profiling platform. Many data processing applications and workflows consisting of multiple modules, which are interconnected by input and output parameters and data, are available free of charge or commercially. Work has been dedicated to construct optimized data analysis pipelines for label-free LC-MS [27, 48], such as Viper [75], OpenMS [76-79], mzMine [80, 81], XPRESS [82], SIEVE, Superhirn [83], Census [84], MapQuant [85], MaxQuant[86], SpecArray [87], MsMetrix [88], PEPPeR [89] or XCMS [90] originally developed for metabolomics but also applicable to analyse proteomics data.

1.1 Data reduction

MS-1 data is three dimensional in nature with retention time, m/z and ion count dimensions. This information is generally stored with succeeding mass spectra storing information in mass – intensity pairs. This raw data is often converted into a two-dimensional regular matrix, with a procedure called meshing, resulting in an intensity matrix, where the columns and rows correspond to a given mass and retention time. Two types of raw mass spectrometry files are provided by the mass spectrometers. Profile data contains all acquired data points, and centroid data is pre-processed by the acquisition software generally with algorithms operating on single MS spectra. Storing data in centroid mode may result in loss of information for certain data processing algorithms, which perform peak detection in both dimensions, but reduces considerably the size of the acquired data. Data processing algorithms, especially those that are written in interpreted, complex high-level programming language such as R or Matlab, generally load all data into the computer memory and are thus limited by the available memory. These algorithms apply data reduction to fit the amount of data to the available memory. This is most frequently done by binning [18, 80, 90] which sums intensities between predefined consecutive and disjoint mass domains. This works well when most of the data points of the Gaussian peaks are within the mass borders of the bin, but leads to fluctuating "saw tooth" type splitting of the peak for centroid data when the bin borders fall in the fluctuation domain of the peak maxima along the consecutive m/z traces. This problem can be avoided by using a two-dimensional Gaussian filter that

smoothens fluctuations in both retention time and mass dimensions out thus avoiding the "saw tooth" splitting of peaks (for details see Figure 3).

Other approaches to reduce the intensity fluctuation of binning were reported recently, however each of them are computationally intensive and results in varying bin widths [91-93]. The quality of the LC-MS data determines the accuracy of feature detection and quantification. Choosing between binning or 2-dimensional Gaussian smoothing of the data has a dramatic effect on quantification when data reduction is applied. Data processing pipelines using programming languages with the possibility to allow user-defined memory management is advised as is the use of streaming to overcome memory limitations in the case of profile data at their original resolution. Streaming is a programming technique which reads and processes only part of the data in one time, and after processing, the results of each parts is written to a continuously growing file. Streaming allows to process large files independently of the available amount of RAM. Data reduction should be avoided if possible due to information loss.

1.2 Noise characterization, feature detection and extraction

A chemical compound with a given charge and isotope distribution is represented as a three-dimensional Gaussian peak in MS-1 and is often denominated as 'feature' in the data processing world. Due to the natural isotope distribution and to the occurrence of multiple charge states, one chemical compound results in multiple Gaussian peaks with the same retention time. These features must, at the first level, be discriminated from noise to determine their main characteristics such as quantity represented by the peak volume, area or height, retention time and mass to charge ratio of the center of the peaks, as well as the extension of the Gaussian peaks in the m/z and retention time dimensions. The second level is the extraction of compound characteristics related to charge state determination and the identification of isotope peak clusters. First level feature characteristics are obtained by all data processing pipelines while extraction of second level characteristics is optional and can be performed at a later stage after matching the same feature of the same compounds in multiple chromatograms. If low resolution mass spectrometry data or considerable data reduction in the mass dimension are used for high-resolution mass spectrometry data, isotope peaks may collapse into a single Gaussian peak or into series of strongly overlapping Gaussian peaks. In this type of data, peak detection algorithms will detect the isotope peaks cluster as one Gaussian peak and provide the average mass of the peak cluster.

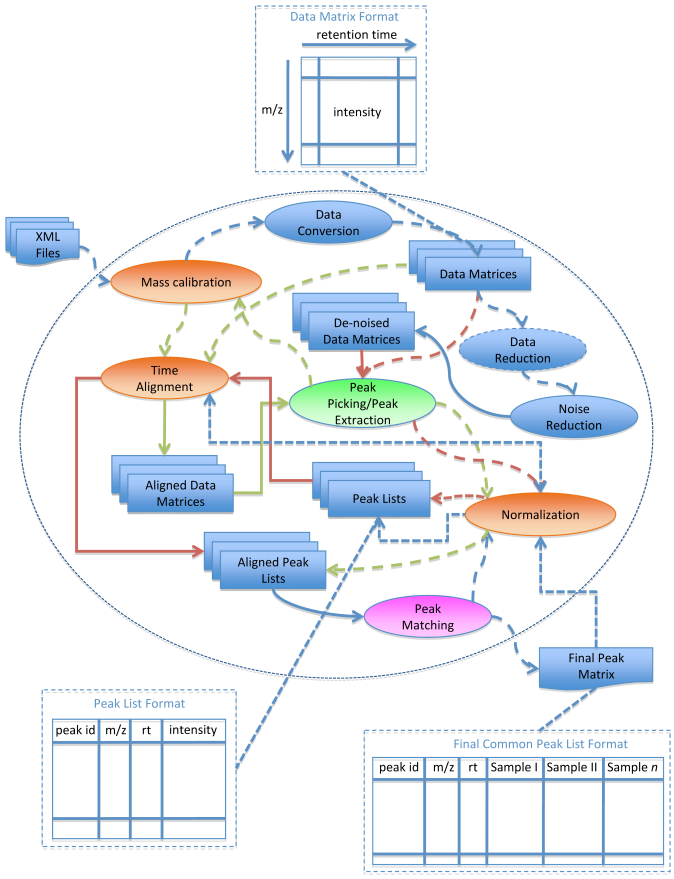

Figure 2. Main modules of quantitative data processing pipelines. Raw data in standard format (mzXML, mzdata.xml or mzML) may be recalibrated for increased mass accuracy and converted to a resampled matrix, where intensity values are in the matrix, rows (or columns) correspond to m/z and columns (or rows) to retention time values (exact values are stored in separate vectors). This matrix may be optionally subjected to data reduction and noise filtering. The obtained data matrix is subjected to peak picking (ellipse in green) and peak quantification providing a peak list containing the most important characteristics of the identified peaks, usually retention time and m/z values of the peak centroid, peak quantity expressed in peak height, area or volume, and optionally peak extension in the m/z and retention time dimensions, as well as charge state and an index that is used to couple peaks of the same isotope cluster or different charge states of the isotope clusters to the same compound. Alignment in the three available dimensions (time alignment, mass calibration and intensity normalization) can be performed either at the peak list or at the raw data file level (ellipses in orange). The aligned and normalized peak lists of different samples are then matched (ellipse in purple), resulting in a quantitative peak matrix containing information about the matched peaks in different samples, where columns (or rows) correspond to different samples and rows (or columns) to different peaks, which are later coupled to identities at the peptide and protein level. This quantitative peak matrix is used for statistical analysis to identify discriminating peaks between predefined classes of samples. The figure indicates the two most common data flows with the order of modules using blue and red arrows. Dashed arrows indicate optional linkage of modules. Another order of the modules and data flows is possible as well the integration of different modules, such as time alignment with peak matching.

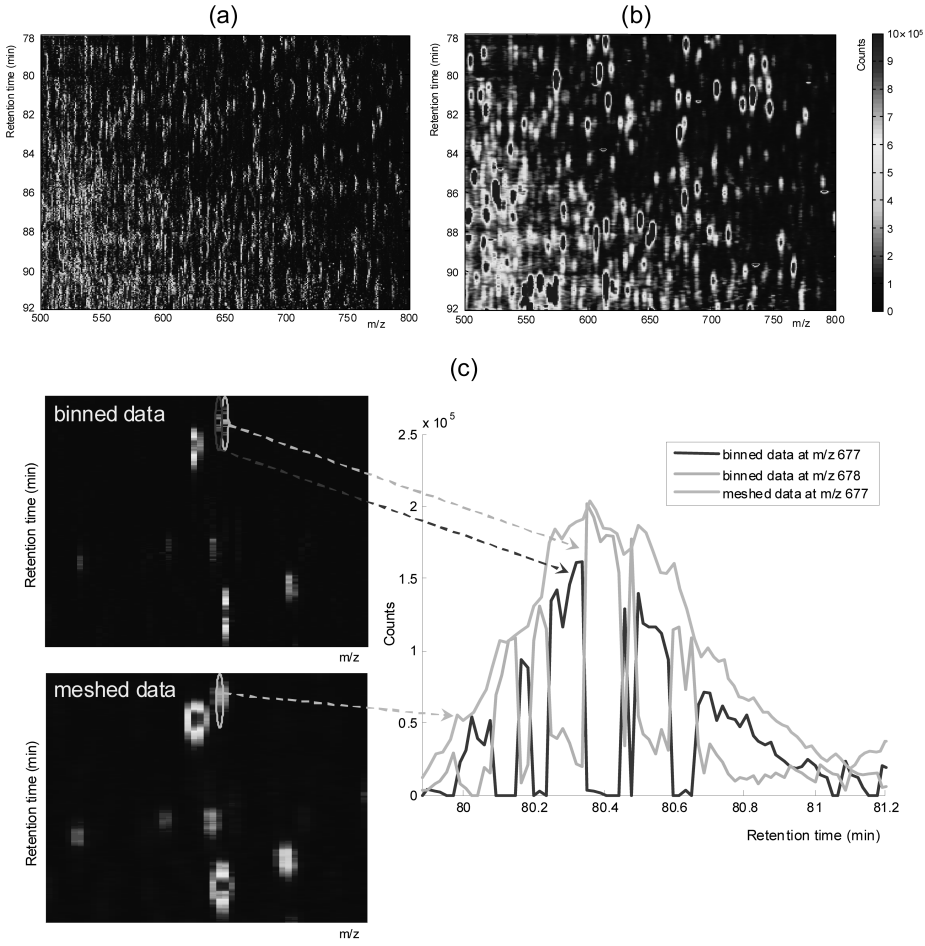

Figure 3. Raw centroided ion trap MS-1 LC-MS image of depleted, trypsin-digested human serum obtained after binning (summing up intensities across 1 amu intervals having borders at fractional decimal of 0.5 m/z for each integer m/z value), (a) and after applying a two-dimensional Gaussian filter using the same degree of data reduction in the m/z dimension as for binning (b). Binning results in noisy data, which leads to a poor feature detection and quantification efficiency, in contrast to the data obtained after Gaussian smoothing. Peak detection becomes extremely difficult in case of fluctuating peak maxima in between mass spectra in two adjacent bins, (c) as represented in a part of an LC-MS image highlighting a peak with extracted ion chromatograms, where the highest intensity of the peak in centroid data fluctuates between the border of the bins. This fluctuation results in a saw tooth peak shape in adjacent extracted ion chromatograms, which will lead to poor performance in case of use of feature detection algorithms recognizing Gaussian peak shapes in individual mass traces.

Noise characterization is important and can be regarded as a part of the peak detection step, which tries to discriminate noise from compound-related features signal. Noise in LC-MS data originates from different sources [94-96]. To discriminate noise from features, it is useful to take the noise model of the different mass analyzers into account

in the peak detection algorithm [97]. Mass analysers and detectors define the background white noise. Another type of noise of chemical origin is called chemical noise. Chemical noise originates from molecule clusters formed during electrospray ionization (e.g. solvent clusters), from chemical contaminations inside the mass spectrometer (e.g. in the ion source), from the chromatographic column or from the ambient air such as polydimethylcyclosiloxanes, phthalate or the plasticizers di-*n*-butyl-phthalate and (di-(2-ethylhexyl)-adipate) [9, 98]. Ion suppression effect [99] distorts the compound-related signals and is dependent on sample composition and therefore on the upstream sample preparation steps. Formation of eluent ion clusters during electrospray ionization and elution of contaminants from the chromatographic columns are highly influenced by the water/acetonitrile ratio of the eluent, which changes gradually during reverse-phase LC-MS. This results in varying chemical background noise in both the retention time and m/z dimensions. Contaminants in the LC eluent or in the ambient air result in stripes of constant m/z across a large part of the retention time axis.

Numerous denoising and baseline (average noise level) subtraction algorithms exist in the literature, such as moving average [18, 100], Savitzky-Golay filters [101] or entropy-based noise reduction [102] to name a few examples. As these algorithms will not be covered in this review, the reader is referred to reviews and publications on this topic [103-108]. It is important to choose baseline removal and denoising algorithms, which do not alter the quantitative information of the peaks. Many feature detection algorithms were developed to discriminate compound-related peaks from noise since the introduction of LC-MS. These algorithms either match the isotope pattern of compounds in the m/z dimension (1D peak picking) or the peak shape based on extracted ion chromatograms (2D peak picking) or on the full 3-dimensional LC-MS data (3D peak picking) to detect Gaussian-shaped peaks. Algorithms that match isotope patterns have the disadvantage that they do not use the full 3-dimensional structure of MS-1 data, but apply peak detection on individual MS spectra, which are subject to a high noise content. VIPER [75], SpecArray [87] and SuperHirn [83] are examples of data processing workflows using peak picking algorithms based on isotope pattern matching. Examples of 2D peak matching algorithms based on extracted ion chromatograms are the M-N rule [18], and the Matched Filtration with Experimental Noise Determination (MEND) algorithms [105]. M-N rules detect an LC-MS feature in extracted ion chromatograms when the intensity exceeds the local baseline N times for M consecutive points. MEND matches Gaussian peak profiles on noise defiltered extracted ion chromatograms. Another example for a 2D peak-shape matching algorithm is based on wavelet decomposition developed by Coppadona *et al.* [106, 107] to remove noise and define the baseline of extracted ion chromatograms. Finally the algorithm uses the difference of baseline and denoised data to detect peaks. Another group of peak detection methods use modified versions of the binning algorithms adapting the bin size to the peak width in mass dimension. These type of algorithms detect peaks in several ways. The first is by locating regions in centroid data with large amount of missing noise in mass dimensions and having high intensity signal with close mass value in consecutive retention time points [93]. Another way of peak detection is by using Kalman filter to extract Pure Ion Chromatograms containing only information on peaks without noise [92]. Other alternative is by detecting peaks containing regions based on analysis of mass variation of centroid intensities in the retention time dimensions and identifying the peaks using

continuous wavelet transformation and optionally Gauss-fitting in the chromatographic domain as implemented in the centWave method [91]. Three-dimensional peak detection methods using shape matching are applied in MapQuant [85], which fits a 3-dimensional Gaussian curve on local maxima. The apLCMS pipeline [104] uses a two-dimensional density kernel function to identify groups of peaks, while MZmine [80, 81], msInspect [109] and LCMS-2D [108] use peak shape in the retention time dimension and the isotopic pattern in the m/z dimension. OpenMS [76-79] uses a three-dimensional wavelet function taking the average of the peptide isotope composition into account by constructing a mixture Gaussian model. There are many other peak detection methods and the reader is directed to a recent review on this topic [110].

To compare the performance of different peak quantification algorithms, the peak picking methods of msInspect and mzMine was compared by analyzing a tryptic digest of a mixture of 48 recombinant proteins resulting in ~800 peptides by FTMS in MS-1 and MS/MS mode with the help of a Receiver Operating Characteristics (ROC) curve. The comparison showed that the isotope pattern matching algorithm of msInspect was superior in performance to mzMine using predefined peak shape template for peak detection [110]. Peak tailing or fronting and saturation of the detector lead to peak splitting for some features. The occurrence of peak splitting depends on the peak detection method and should be evaluated for each algorithm using different types of data. Algorithm developed by Groot *et al.* [111] uses K-mean clustering to correct for split peaks and to correct peaks that were incorrectly aligned in the retention time dimension. Other quality control criteria applicable only to high resolution data such as Orbitrap and FTMS uses mass deviance to assess if a detected compound correspond to real peptides [112]. Mass deviance is the difference of the decimal fraction of the monoisotopic peak of the detected compound and the nearest theoretical tryptic peptide. Overlaying on a part of a 2 or 3 dimensional MS-1 raw or pre-processed data with the location and extent of detected features as it is possible in OpenMS framework enables to assess visually the accuracy of peak picking method [76, 78, 79].

1.3 LC-MS Map Alignment

Three-dimensional MS-1 LC-MS chromatograms are prone to nonlinear shifts in all of the three dimensions. In the mass dimension, alignment is based on proper calibration of the mass analyzer preferably with internal standards. In the intensity dimension, normalization may be used, and in the retention time dimension time alignment is necessary. Mass calibration should provide alignment to the exact mass. Intensity normalisation could be relative to compare relative intensity of the peptides and proteins, but may provide exact alignment if absolute quantification is required. Retention time alignment is relative, in spite of the fact that retention time indices may be used for identification [113]. After successful alignment of all LC-MS chromatograms common peaks in different chromatograms are matched and their relative or absolute quantities are reported in form of a matrix that is amenable to statistical analysis. In this matrix, rows (or columns) correspond to samples and columns (or rows) to features or peptide peak identities. Alignment of LC-MS images can be performed with different goals in mind depending on the experimental design, such as to transfer peak identity

information from separate MS/MS data sets to MS-1, or to combine data from several chromatograms corresponding to the fractions of a 2D-LC-MS analysis of a single sample.

1.3.1 Mass calibration

The m/z dimension is the most stable dimension toward shifts. The absence of shifts does, however, not mean that the measured values are accurate. This requires calibration of the mass spectrometer, preferably with internal standards that are present in each spectrum. Mass analyzers are, however, measuring instruments that are prone to small nonlinear shifts requiring automated algorithms to compensate for inaccurate mass calibration or to enhance mass accuracy, especially for high resolution mass spectrometers. Ions of chemical background noise originating from eluents or from the ambient air such as polysiloxanes or continuously added calibration standards can be used for mass calibration and to increase mass accuracy. A polynomial mass calibration function was used by Scheltema *et al*. [114] to increase mass accuracy of metabolites measured with an Orbitrap mass spectrometer. The algorithm improved mass accuracy from 1-2 ppm to 0.21 ppm using background ions, such as polysiloxanes, as internal standards. Haas *et al*. [115] used polydimethylcyclosiloxanes to enhance mass accuracy of MS/MS spectra and reported a higher identification rate of peptides. Another strategy involves the use of already identified peptides as calibration standards, a strategy that was successfully applied to improve peptide identification based on MS/MS spectra [116]. An interesting approach developed by Dijkstra *et al*. [117] superimposes isotope clusters of the same peptide at different charge states in SELDI-TOF-MS spectra to improve mass accuracy. This approach can be easily adapted to spectra obtained with other instruments and electrospray ionization, as multiple charging of peptides is a common phenomenon.

1.3.2 Intensity normalization

High throughput LC-MS data come with nonlinear and systematic bias in recorded peptide ion intensity, affected mostly by differences in injected sample amount, differences or drifts in ionization efficiency, differences in ion transmission efficiency or detector saturation, and carryover between LC runs. The resulting bias should be corrected in order to enhance statistical classification accuracy. Systematic bias due to a difference in injected sample amount should be minimized, e.g. by determining the injected amount with a total protein assay or by taking the area under the curve of the UV trace of a previous analysis into account. Sometimes normalization can be based on a single factor such as the average or median abundance of peptides derived from so-called 'housekeeping proteins' or other compounds that are known not to be affected by the investigated disease or sample dilution factor, such as creatinin in urine samples [118]. However, all intensity normalization approaches have drawbacks, *e.g.* normalization based on housekeeping proteins should not be applied for samples changing in constitution of these proteins considerably. Remaining nonlinear bias can be removed using normalization methods applied in microarray analysis, [119] e.g. by performing a nonlinear regression of matched peak intensities in two samples with the same or close

composition. Various normalization methods were developed and assessed for label-free quantification using LC-MS [101, 103, 120-122]. Linear regression can be applied to a bias that has a linear pattern across analyses such as sample carryover in the trapping column. Nonlinear bias caused *e.g.* by detector saturation can be resolved by non-linear or local regression techniques, and quantile normalization may be used to transform peptide quantity distribution of all samples to the same distribution using quantile plots. Callister *et al.* [121] compared different normalization methods using LC-FTICR-MS data sets and concluded that global or linear regression worked best in most cases when applied iteratively. A recent study by Kultima *et al.* [120] compared 10 different normalization methods using data sets from mouse, rat and quail that were analyzed by a nano-LC-MS system coupled either to a Q-TOF or an LTQ mass spectrometer. Karpievitch *et al.* [122] developed a normalization method based on singular value decomposition to remove systematic and nonlinear bias to avoid over fitting by dimension reduction for label-free LC-MS proteomics samples. In general, normalization algorithms use matched peak matrices. Therefore normalization procedures are implemented after peak matching and prior to the statistical analysis.

1.3.3 Time alignment

The retention time of compounds is subjected to considerable non-linear shifts between LC-MS experiments and requires particular attention and more sophisticated alignment algorithms than the two other domains. Complex proteomics samples, such as body fluids like serum and urine, contain several tens of thousands of peptides, so that even small retention time shifts may result in serious peak mismatching, if peak matching is only based on the retention time and m/z coordinates across multiple samples. Retention time shifts are due to parameters of liquid chromatography that are hard to control, such as small changes in eluent composition, pH, column ageing or temperature changes and have a highly non-linear behavior, especially when the combination of a trapping and separating column system is used. Accurate algorithms to correct retention time shifts is one of the most critical points of data processing to provide accurately matched peaks and quantitative data that are suitable for statistical analysis.

The goal of time alignment is to find the greatest overlap between the same peaks in different chromatograms and to provide a retention time transformation function, which can be used either to change the retention times of all peaks in a peak list, or to change the associated retention time of mass scans in the raw data. The major difference between time alignments methods using MS-1 data is how many data dimensions they use in their benefit function to drive the time alignment procedure. Earlier developments considered only 1-dimensional data next to the retention time dimension in their benefit function (e.g. the TIC or BPC) [123-131]. Recent algorithms use two-dimensional profiles that take the separation of compounds in the mass and the retention time dimension into account. The latter approaches provide more accurate time alignment of highly complex LC-MS 'omics' data. Two-dimensional alignment algorithms differ in terms of whether they use the raw data, pre-processed data obtained after noise filtering and data reduction [95, 96, 101, 127, 128, 132-134] or peak lists after the peak detection step [77, 97, 135].

A large number of algorithms were developed to define the optimal search space for non-linear retention time correction, such as Correlation Optimized Warping [96, 97, 123, 127, 129, 130, 136], Parametric Time Warping [95, 125], Dynamic Time Warping [95, 101, 126, 128, 130-134, 137], a geometric approach based on pose clustering [77, 138], Loess regression on matched compound pairs [135], the Continuous Profile Model combined with a Hidden Markov Model [139] to list a few. Time alignment based on DDA MS/MS data use the correlation between MS/MS information of the same compound [89, 140]. Other types of algorithms create retention time and mass tags by normalization of retention time and accurate mass. These tags are subsequently used to align multiple LC-MS data sets in both the m/z and retention time dimensions or through comparison with a database [137].

Generally time alignment is performed by selecting one chromatogram as reference and aligning all others to that reference pair wise. This approach requires the *a priori* selection of a reference chromatogram and must assess how selecting different reference chromatograms affects the quality of time alignment. Robust time alignment methods should not depend on the choice of the reference chromatogram [95, 96]. The Continuous Profile Model developed by Listgarten *et al.* [139] does not use a reference chromatogram, but performs the alignment of all chromatograms in one step. The performance of different time alignment algorithms depends on many parameters, such as the number of common peaks shared between chromatograms, the complexity of the samples, the compound distribution in retention time – m/z space, the compound concentration variability and noise distribution. A comparison of different algorithms with different characteristics shows that time alignment methods that take the three-dimensional nature of MS-1 data into account perform better for complex proteomics samples with large compound concentration variability [95-97].

In most studies the same chromatographic columns and strict standard operating procedures are applied in order to lower analytical variability. It is rare that the elution order of compounds changes under these conditions and a monotonic time alignment function is appropriate [141]. However the elution order may change during extensive studies over a long period of time or when different types of columns are used (e.g. when using different types of *n*-octadecyl bonded silica reverse phase stationary phases) [142]. It is also known that the pH of the eluent has a dramatic influence on the selectivity of reverse phase (RP) columns [143, 144], and this can lead to a changing elution order when analysing complex proteomics samples. Inversion of the elution order of peptides or metabolites is not commonly reported and thus probably not recognized. It is important in the future to explore this phenomenon in greater detail, especially when different types of LC RP-C18 columns are used within one study or when large-scale studies are performed in different laboratories. If changes in peak elution order were frequently observed, this would require novel time alignment algorithms, which can adequately with peak elution order changes. For further reading on time alignment, the reader is referred to specialized reviews [141, 145] and articles presenting results from performance comparisons of different time alignment methods [95, 130, 138].

The quality of time alignment is generally evaluated by visualization of the entire chromatogram, or by visualization of common peaks to assess the local time alignment accuracy. Co-injected standard peptides or peptides derived from highly abundant housekeeping proteins can also be used for this purpose. Comparing the quality of different time alignment algorithms in such a way is a daunting task and

visualization of entire chromatograms does not always allow a proper quality control of the time alignment results. A quality assessment method based on the sum of the overlapping peak area between pairs of chromatograms provides a global readout of alignment quality and permits comparison of the relative performance of various types of algorithms [95, 96]. A Similarity score calculated after time alignment beside for assessing the quality of time alignment, may also be used to assess eventual bias in experimental design and to detect whether there is a systematic difference between sample replicates or injection order [146].

1.4 Peak Matching

The peak matching process identifies common peaks in different chromatograms either based on proximity of aligned retention time and mass or peptide identity based on MS/MS data when this information is linked to the peak list. Numerous clustering algorithms have been applied to match peaks such as K-mean, hierarchical or pose clustering [103, 145, 147]. The procedure provides a list of clusters of the same compounds in different chromatograms, from which a quantitative feature matrix is constructed. The matrix contains a quantitative measure of the feature and rows (or columns) that correspond to the samples and columns (or rows) to the features (e.g. peptide ions) characterized by retention time and m/z value. Features in this matrix can be further processed by deisotoping and by integration of different charge states of the same compound resulting in a quantitative peptide matrix, which further may be combined with identification results. Deisotoping and decharging can also be performed at peak list level prior to the peak matching procedure. MS-1 data processing pipelines must assign a quantitative value to peaks that do not have a correspondence in all chromatograms. Some pipelines extract noise at the corresponding location, while others filter out peaks found only in a minority of samples to avoid bias for single and rare events in the subsequent statistical analysis.

1.5 Peptide and protein identification

Peptide and protein identifications are generally performed using information from MS/MS spectra. Before identification, MS/MS spectra are filtered to remove noise and the cleaned spectra are used for the identification process. The most widely used protein identification approach is database search, where lists of peaks in MS/MS spectra are compared with molecule fragments obtained by *in-silico* fragmentation of sequences stored in the database [148-150]. This comparison may be performed by calculating a similarity score between the *in-silico* fragments and the measured fragments, and the peptide with the best match receives the highest score. A threshold is used to limit the number of false positive identifications while at the same time avoiding to penalize true positives. It is thus noteworthy that so-called identified peptides and proteins always contain a chance that they are false positives. This matching approach is used by the Sequest algorithm [151]. Interpretative models use the assumption that MS/MS spectra consist of a continuous series of fragment ions that can be interpreted as a partial short amino acid sequence tag of the intact peptide. PeptideSearch [152], MS-Seq [153] and GutenTag [154] use this strategy. Stochastic model-based algorithms use probability

estimates for peptide fragmentation and subsequent predictions of the resulting mass spectra that are compared with the measured MS/MS spectra. SCOPE [155] and Olav [156] (the basis of the Phenyx search engine) are examples of this category. Finally programs such as Mascot [157] and the open source OMSSA [158] apply statistical and probabilistic models using empirically generated ion probabilities of peptide sequences stored in the database. Spectral library search algorithms represent another category of peptide identification algorithms. These algorithms compare noise-filtered MS/MS spectra with databases containing high quality, experimental MS/MS spectra using similarity scores [159, 160]. Clustering MS/MS spectra of the same peptide can enhance the probability for successful peptide and protein identifications significantly while at the same time decreasing the number of spectra that are sent for database search by one order of magnitude [161]. Database search and spectral library search algorithms have limited capability to identify peptides with PTMs, since the peptide with a given PTM should be either present in the database or must be defined by the user prior to the search. Open modification search programs such as Popitam [162] and Inspect [163] use MS/MS spectra of already identified peptides and allow unexpected mass shifts in the fragmentation pattern of the peptides due to PTMs. Open modification search algorithms are, however, computationally more intensive than database search algorithms and therefore they generally use a limited number of peptide sequences for identification. *De novo* sequencing algorithms such as PEAKS [164, 165], PepNovo [166], EigenMS[167], Lutefisk [168], Sherenga [169], MSNovo [170], PILOT [171], NovoHMM [172], and AUDENS [173] use only information from the experimentally acquired MS/MS spectra and basic constants such as the mass of the amino acids to elucidate the most probable sequence of the fragmented peptide.

A simple peptide identification method using label-free MS-1 data exploits so-called accurate mass and time tag (AMT) information that is calculated from accurate mass and retention times with or without normalization to match pre-identified peptide sequences in a database to the newly acquired data [113, 174, 175]. This identification strategy has the advantage to perform MS/MS based peptide and protein identifications on pooled or representative samples using time-consuming profiling techniques with a large peak capacity (e.g. 2D-LC-MS with DDA MS/MS data acquisition) followed by the quantitative analysis of a large number of samples with faster LC-MS platforms operating in MS-1 mode, which cover a larger measured concentration range. Basically this technique can be considered as a generic peptide and protein identification transfer system that uses mass and retention time information for the matching, and has the disadvantage that the identification transfer may be sensitive to the LC parameters and that high-resolution mass spectrometers with high mass accuracy (FTMS or Orbitrap) are required to reduce the chance for incorrect matching to acceptable rate [174, 175]. In most case normalization of the retention time is performed through regression of the observed and predicted retention times using training data sets and a neural network for retention time prediction. The databases containing peptide identifications with AMT tags are generally obtained from different analyses by generating reference maps [113, 176].

To decrease false positive identifications, Scaffold [177] combines identification results of different database search programs such as Sequest [151], Mascot [157] or X! Tandem [178] (a version performing parallelized processing is called X!! Tandem [179]) and calculates a composite probability score, providing more reliable protein identification compared to

the single score of one program. A protein identification score is constructed from identified peptide scores and relates to the probability that a given identification is a true positive. With its combined scores, Scaffold also provides a more reliable grouping of peptides for protein identification. The probability for false positives can be further decreased by comparing the measured and predicted retention times of the identified peptides. Retention time prediction algorithms use statistical methods based on quantitative structure–retention time relationships [180], which are in turn based on a large number of molecular descriptors or training data sets and regression methods taking the amino acid composition of the peptide into account [181, 182].

Database search algorithms are biased to the proteins present in the database and are poor to detect splice variants and proteins with PTMs. In order to be able to identify proteins with splice variants and PTMs, the Swisspit [183, 184] workflow combines the results of identifications obtained with the Phenyx [156] and X! Tandem [178] database search type algorithms with the Popitam and Inspect open modification search type algorithms. This is a combination of first assigning identifications using database search programs and subsequently submitting unassigned MS/MS spectra to open modification search algorithms with restriction to use peptide sequences identified previously by both database search tools. The Swisspit [183, 184] workflow resulted in a higher identification rate of 77% for small well-annotated PTM-rich data sets compared to 21% obtained with the combination of the two database search programs only.

Another ingenious approach uses spectral network analysis for peptide identification [185]. This method finds pairs of MS/MS spectra that differ only in one modification or amino acid by searching for corresponding *b*- and *y*-type fragment ions. From multiple paired spectra, a network is constructed and is used to propagate peptide identification from peptide without PTM to the same peptide with different numbers of PTMs or amino acids changes. Spectral network analysis was further adopted to include data from multistage MS/MS such as MS^3 or MS^4 in the interpretation [186]. Spectral dictionaries extend the sequence tag approach by generating sets of full-length peptide *de novo* reconstructions. These spectral dictionaries are then searched in a database equipped with hash table or suffix tree providing a fast identification algorithm, with high true positive identification rates [187].

Since database search algorithms provide always a list of identified peptides and proteins with given scores, it is important to test the statistical significance of the obtained score against a decoy database containing incorrect protein sequences obtained, for example, by reversing or randomizing existing protein sequences [188-191]. The presence of highly homologous proteins represents an actual challenge for the protein identification software, therefore in case of lists of proteins with high sequence homology the results should be taken with precaution specially, when only peptides shared with other proteins are identified. The reader is referred to dedicated books and reviews for further reading on the peptide and protein identification algorithms [148, 149, 192-194], quality control methods [195] and influence of parameters affecting the quality of MS/MS [196].

2 STATISTICAL ANALYSIS AND VALIDATION

2.1 Coupling feature quantification with peptide and protein identity

Quantitative feature matrices should be first transformed to quantitative compound matrices by summing up the quantity of isotope clusters and different charge states of the same compound (a one signal per compound matrix). Peptide quantity should be further matched with compound identity at the peptide and ultimately at the protein level. Regarding protein quantity, different methods can be used starting from summing up the intensity of constituting peptides to taking the sum of the three most abundant peptides of each protein [55]. Mapping peptide identity to the quantity of extracted features involves several steps. Precursor ion mass of a given charge state should be matched to the isotope cluster of the same charge and the corresponding quantity (e.g. represented by the sum of all isotopomer peak heights or the peak height of the monoisotopic peak only) of the isotope cluster should be combined with the quantity of the other isotope clusters of the same peptide with different charge states. This step is followed by the determination of all identified peptides constituting individual proteins. The peptide centric nature of the shotgun proteomics approach makes quantification of original protein mixtures challenging in the presence of multiple proteins with high sequence homology, truncated protein forms, proteins having different PTMs or multiple splice variant [6, 7]. For accurate protein quantification, either with MS-1 or spectral counting methods, precise peptide and protein identification including detection of all protein variants is necessary, because the identified peptides will provide the list of peptides unique for the protein and peptides that are shared between several proteins, which will allow accurate quantification of all protein variants. Recently Zhang *et al.* [197] evaluated different strategies for spectral counting quantification and found that the most accurate quantification were obtained by adding the corresponding molar proportion of the spectral counts of peptides shared between different proteins to the spectral counts of unique peptides for the protein. The identification of exact protein forms is also important for the development of accurate targeted MRM assays, while the presence of peptides shared between proteins could bias considerably the measured protein quantities [6, 7, 60]. A recent review by Podwojski *et al.* [198] deals with this problem in detail.

Annotated quantitative peptide matrices can be obtained through other methods [199] than those based on MS-1 data, such as the already mentioned MRM-based methods, spectral counting algorithms and by quantifying spot intensity in two-dimensional gel electrophoresis coupled with peptide fingerprinting or LC-MS/MS identification [200-202]. Immunochemical techniques based on antibody arrays are especially interesting for the targeted profiling and validation of proteins in complex biological samples [203-205]. Quantification methods either provide absolute or relative protein quantities or other type of bio descriptors if the goal of the analysis is to compare biological patterns for sample classification [206].

2.2 Feature selection/transformation methods

The main application of statistical analysis, also called post-processing methods, is to find peptides and proteins that discriminate between different groups of pre-classified samples. Discriminating peptides or proteins are selected from the common peaks after data processing, which means that the validity of the ultimate statistical result depends on the quality of data processing. Statistical analysis of quantitative proteomics experiments suffer from the high dimensionality given by the large number of identified peptides and proteins accompanied by a much lower sample size. Due to this characteristic, processed proteomics data are often referred as megavariate data [207, 208] leading to a High Dimensionality Small Sample size (HDSS) problem [147]. HDSS is the main reason why most of the widely used classification methods such as linear discriminant analysis cannot be directly applied to analyze quantitative proteomics data sets. In data sets with HDSS properties, a large number of compounds may be found to differ significantly between predefined classes of samples using, *e.g.* the univariate *t*-test, but they may not be confirmed in other independent set of samples. Disease-related changes generally affect a small portion of proteins and peptides in living organisms, which represent the truly discriminating molecules between predefined groups of samples, and which stay true when measuring new sample sets. To find a small number of truly discriminating proteins among a very large number of other non disease-related proteins in data sets with HDSS, it is necessary to either use statistical methods that are insensitive to uninformative features (noise) or to reduce the number of features (dimensions) prior to the actual statistical analysis to a number that does not exceed the number of samples (independent observations). Only very few methods such as Support Vector Machines [209-211] or Learning Vector Quantizations [212, 213] claim to be insensitive to a large amount of noise peaks contained in HDSS data sets and generally do not require upfront feature selection. Other methods require dimension reduction, which can be performed either by removing uninformative peaks and selecting statistically relevant discriminative peaks (so-called feature selection) or to perform data transformation to accentuate class differences. Feature selection methods are most widely used [103, 214-217], since these methods are not only helpful to overcome the HDSS problem, but also to provide a list of discriminatory peaks. Selected features corresponding to a limited number of biomarker candidates must be validated by measuring a larger sample set with fast and targeted analytical methods such as LC-MS/MS in the MRM mode [3]. The results can also be used as input for databases and algorithms to link them to biochemical, secretion, molecular interaction or signaling pathways that may be involved in disease-related biological processes.

Feature selection can be performed in a supervised manner using univariate or multivariate selection algorithms. Univariate methods assume that features are mutually independent, so that each feature is evaluated individually based on its individual relevance to discriminate between predefined classes of samples. The simplest method for feature selection is the univariate Student's *t*-test, which must be corrected for multiple testing. Multivariate feature selection methods take the interdependency between features into account, when evaluating the individual strength or rank of a given feature. Collective assessment and selection of variable subsets is another type of feature selection. This method selects a feature subset by evaluating all possible correlations or other forms of dependencies between features.

Since the number of subsets increases exponentially with the number of features, it becomes an exhaustive task to evaluate all possible subset combinations in the feature space. Therefore, most collective feature selection methods are based on heuristic search strategies, such as forward selection and backward elimination [218]. Forward selection methods start with an empty feature subset and add features step-by-step to maximize a predefined scoring function. The procedure is stopped when newly selected features have a small contribution to the value of the scoring function. Backward elimination starts from the full feature set and eliminates features until a given scoring function reaches its maximum. An example of such an approach is the Reduce Feature Elimination method that can be combined with a classifier such as a Support Vector Machine [218]. Feature transformation methods construct new features from the original features while maintaining the initial data structure as accurately as possible. Typical feature transformation methods create a supervised or unsupervised mapping function that changes the initial feature space into a transformed variable space. One of the most widely used methods for feature transformation are Principal Component Analysis (PCA), Fourier and the wavelet transformations. The most popular feature transformation methods coupled to statistical classification are Principal Component Linear Discriminant analysis and Partial Least Square Linear Discriminant Analysis. Many feature selection methods are available and it is difficult to predict, which of them perform the best with respect to the others. Since many of the features in biological samples are correlated, in context, collective feature selection and feature transformation methods taking account for this correlation are preferable.

2.3 *Classification methods and statistical validation*

Figure 4 shows the main modules of statistical data analysis and validation [215, 217]. Feature selection and statistical analysis form the core, which is surrounded by different stages of validation layers to ensure that the resulting classification models are robust with respect to new sample sets. From the core modules, only the classification module is compulsory while feature selection and ultimately biomarker selection are optional if the classification method is not sensitive to HDSS and if sample classification is the only goal without identifying the underling molecular determinants. The first validation layer serves to select the optimal classification model (module 4) and the second layer measures classifier performance (module 5) by providing an error rate when classifying new samples, that were not used for building and selection of the classification model. Finally the relevance of the discrimination of the model is determined by comparing its performance to models obtained by chance using *e.g.* permutation tests (module 6). Permutation tests randomly reassign sample group labels thus generating random models and the performance of large number of random models is compared to the performance of the model obtained with the correct sample labels [217].

Generally validation modules 4 and 5 are based on a double cross-validation strategy, with model selection occurring in the inner loop and classification performance being determined in the outer loop. Double cross validation strategies provide an unbiased way of evaluating model selection criteria and classification performance. This is achieved by dividing samples in each group into a training set, which is used for

model building, and a test set that is used for performance measurement in the inner and outer loops, respectively. Another way to assess model performance is to determine the sensitivity and specificity of the statistical model by calculating the so-called ROC curves [215, 219]. Current developments in statistical analysis focus on a combination of several methods resulting in ensemble classifiers. The reader is referred to more comprehensive reviews for details about feature reduction, dimensional reduction, statistical analysis and validation methods [147, 217, 220].

Most biomarker discovery studies are performed using samples from different groups, which are obtained from different individuals (e.g. patients or animals) not related to each other (cross-sectional study design). A study design where samples from the same individual are analyzed at different time points (longitudinal study design) is able to lower the biological variability. It is possible to further decrease the effect of biological variability by matching the different clinical parameters such as age, sex, smoking habits or life style. In this case statistical analysis should take the relation between samples into account by using adapted methods (*e.g.* time series analysis) to correlate compound concentration to time or other parameters such as drug dosage or disease development status [221-224].

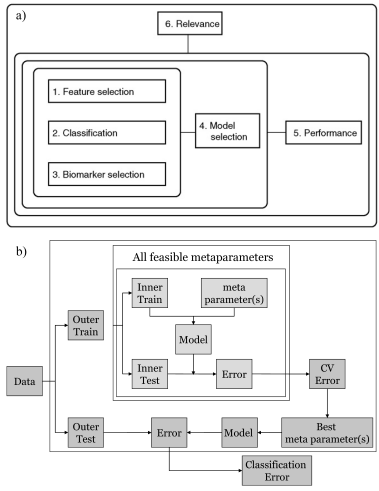

Figure 4. Main steps of the statistical analysis and validation strategy for proteomics data suffering from the HDSS problem. Panel a) gives a schematic overview of the modules for supervised statistical analysis and validation. The core (modules 1-3) represents the statistical analysis modules for feature selection (module 1; optional), classification (module 2) and biomarker selection (module 3; optional). The outer part represents modules for statistical validation comprised of a model selection module (4), a module to assess performance of the classifier (5) and a module assessing the relevance of the selected biomarkers by permutation tests (6). Panel b) gives an overview over the generic double cross validation strategy to measure the performance of the feature selection-classifier modules (modules 4-6). Figures taken with reprint permission from Smit *et al.* [215].

Quantitative peak matrices can be used for different purposes than identifying class specific discriminating compounds. Figure 5a shows the histogram of the relative standard deviation (RSD) of the quantity of all compounds in a given data set, where variability was only subjected to analytical variance due to the use of two different LC-MS analysis platforms. The histograms show that the two platforms perform equally well to quantify peptides in serum samples depleted of the six most abundant proteins [100]. The aligned peak matrix contains also information about the global concentration variability of compounds in the different data sets. This comparison can be performed similarly using histograms of the RSD as presented in Figure 5b for three different types of sample sets [96].

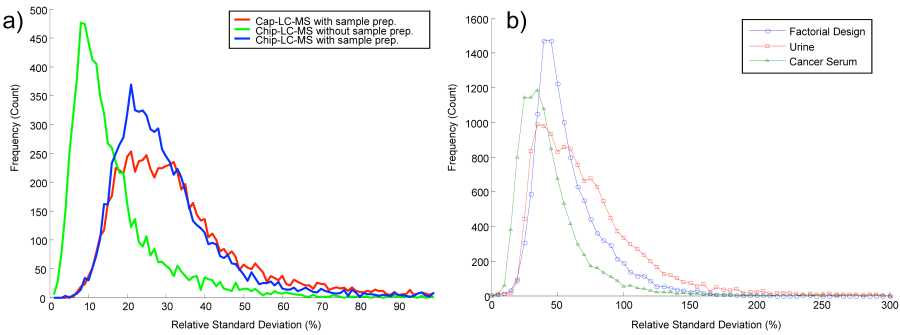

Figure 5. Histograms of the relative standard deviation (RSD) of compound concentrations calculated based on results of the corresponding quantitative peak matrices. a) Assessment of two LC-MS profiling platforms used for comparative proteomics studies. The Chip-LC-MS platform was equipped with a reverse phase nano-LC column (75 µm internal diameter) integrated in a microfluidic device and coupled to the mass spectrometer via an electrospray interface, while the Cap-LC-MS platform used a 1 mm internal diameter reverse phase column coupled to the mass spectrometer via an electrospray interface using a nebulisation gas (ionspray). The histograms show that the Chip-LC-MS (in blue) and the Cap-LC-MS (in red) platforms result in similar compound concentration variability and can thus be considered as equivalent quantitative profiling platforms. In both cases 10 serum samples from the same patient were depleted of the 6 most abundant proteins and underwent individual sample preparation procedures. These data thus contains only the analytical variability. Five injections of the same sample in Chip-LC-MS (green) resulted in histograms with a lower RSD indicating that most of the analytical variability is not caused by the Chip-LC-MS profiling platform itself but originates from the sample preparation steps (depletion or trypsin digestion). b) Assessment and comparison of compound concentration variability after accurate data processing of different types of body fluid analyses from ongoing biomarker research projects. Twenty chromatograms of trypsin-digested human serum samples obtained from 10 different patients at two time points and depleted of the 6 most abundant proteins give a narrower RSD histogram with respect to compound concentration (green histogram) than the same type of samples obtained from one patient and subjected to an experimental design study (19 chromatograms) obtained by varying pre-analytical parameters (blue histogram). Fifty acid-precipitated human urine samples result in the widest RSD distribution (red histogram), which may be explained by the fact that serum is a well-regulated body fluid while urine is excreted and thus devoid of homeostatic control. Figures taken with reprint permission from Horvatovich *et al.* [100] and Christin *et al.* [96].

The figure shows that there are large differences in concentration variability between sample sets of different origin, such as acid-precipitated urine, serum depleted of the 6 most abundant proteins and a serum sample from one patient that was subjected to varied pre-analytical factors in a factorial design study. While the compounds of the first two data sets were subjected to biological and analytical variability in addition to errors during data processing, compounds of the last data set were only subjected to analytical variability and data processing errors. The histograms indicate that well-regulated body fluids, such as serum, show less concentration variability than the same type of sample measured with different pre-analytical factors in a factorial design study. An excreted, non-regulated body fluid, such as urine, shows the largest concentration variability. Finally, quantitative peak matrices can be used to evaluate and compare the quality of different data processing pipelines using data sets containing compounds that were added in known concentrations by spiking. Grossmann *et al.* [47] compared the quantification performance of two relative spectral count methods (emPAI and APEX), an absolute protein quantification method using the abundances of the three most abundant peptides developed by Silva *et al.* [55] and four different protein quantification methods using Superhirn (MS-1 based quantification) by assessing the robustness and dynamic range of the spiked-in protein as well other non altered proteins detected in the mixture of spiked yeast samples. The protein quantification method of Silva *et al.* [55] with own implementation or using Superhirn [83] provided the best performance compared to emPAI or APEX methods, however their results should be taken carefully as they have used one bovine protein (Fetuin-A of 38.4 kDA) for spiking, which has a very different composition (low homology) than the yeast proteome. One other recent paper compares the quantification performance of two commercial label-free LC-MS data processing softwares (Eluciator and Progenesis) based on spiked samples, and reported considerable differences in performance [225].

3 CONCLUSIONS

Improvement and development of new data processing pipelines and individual modules will continue into the future as mass spectrometry-based molecular profiling is gaining momentum in life science research and developments of new mass spectrometers and new sample preparation methods are on the agenda of numerous research groups and companies. In current bioinformatics literature, there is an increasing emphasis on the development of algorithms, which assess and compare the performance of data processing methods. These algorithms also provide substantial support for parameter optimization and troubleshooting of algorithms. Performance assessment and quality control can be only performed with high quality standard data, where compound composition and quantity are precisely known [226]. To assess the performance of protein identification workflows, which report enhanced performance to detect known and unknown PTMs, there is a need of open access well annotated MS/MS data [183]. Similarly, for the purpose of assessing quantitative label-free data processing workflows, it is necessary to provide reference data from different biological origins such as urine, serum treated with different depletion techniques or different cell lines spiked with known compounds in known concentrations. Development and access to standardized samples such as the recently introduced yeast standard sample help this

procedure [226]. Well-characterized data sets are also helping to evaluate the performance of statistical analysis and validation strategies. Raw and processed data simulation softwares could, in certain circumstances, replace real data sets or create datas ets with particular properties, which are difficult to obtain experimentally, and which may reduce assessment time considerably. For this reason, well characterized and documented data sets stored in free access database such as Human Proteinpedia [227-229], and the development of accurate data simulation tools will, in the future, enhance the comparison and assessment of different modules and complete data processing workflows. An example for quality assessment and parameter optimization of time alignment algorithms using well-defined samples is provided by Peters *et al.* in a recent publication [230].

Another trend is to develop data processing solutions to integrate highly diverse data such as data obtained with different instruments or in different laboratories. An example of such diverse data, which can not be processed with the actual data processing pipelines, is presented in Figure 6 showing a representation of two raw LC-MS data sets that were obtained form the same serum sample with the same ion trap mass spectrometer, but using two different ionization methods and two different HPLC techniques. While the two samples contain exactly the same compounds with same concentrations and measured dynamic concentration range, the differences caused by different chromatographic methods and ionization modes result in different peak distributions in the retention time and m/z space. Current LC-MS data processing workflows are not able to accurately combine these types of data sets, leaving this challenge for future developments.

Newly developed or enhanced algorithms are emerging rapidly within bioinformatics research groups. However these new algorithms, with a high potential to ameliorate information extraction accuracy from raw data and biological knowledge discovery, are not used by the majority of the data producing, application-oriented proteomics laboratories. The main reason for the low penetration of new bioinformatics solutions is that mass spectrometer vendors generally provide user-friendly data processing and evaluation pipelines supported by training sessions favoring the application of their own software packages, even if the performance of these softwares is not assessed and compared with others. The newly developed algorithms, even if the open source program code is available, require on the other hand extensive bioinformatics expertise, which is not present in most data-producing proteomics laboratories. In order to allow a breakthrough for the widespread application of newly developed algorithms and software tools, it is necessary to develop infrastructure programs, which provide data processing services using integrated tools with access to high-capacity, parallel computing facilities, such as large local clusters or grid. Indeed biologists planning proteomics or in general life science experiments to answer relevant biological questions may work more efficiently if they have to concentrate only on the experimental design of the biological study, on production of high quality data and interpretation of the obtained data using easy to use, user friendly data processing services. To facilitate data interpretation, the complexity of the software and hardware operations should be hidden, and the end user should be only exposed to raw data management, parameter setting of data processing and other bioinformatics tools, to monitor the data processing status and to visualize the processed data in a user-friendly

way. An example for such a framework software is Galaxy [231, 232] or Genepattern [233, 234], which is extensively used to analyse new generation DNA sequencing data. The framework should make the integration of new bioinformatics tool easy and allow to modify complex data processing workflows to adapt to the large diversity of mass spectrometers and sample preparation methods that generate highly diverse data. A key element for the efficient and easy integration of diverse bioinformatics tools in such a software framework is to use a standard format, which serves to interconnect the input and output files of the integrated tools.

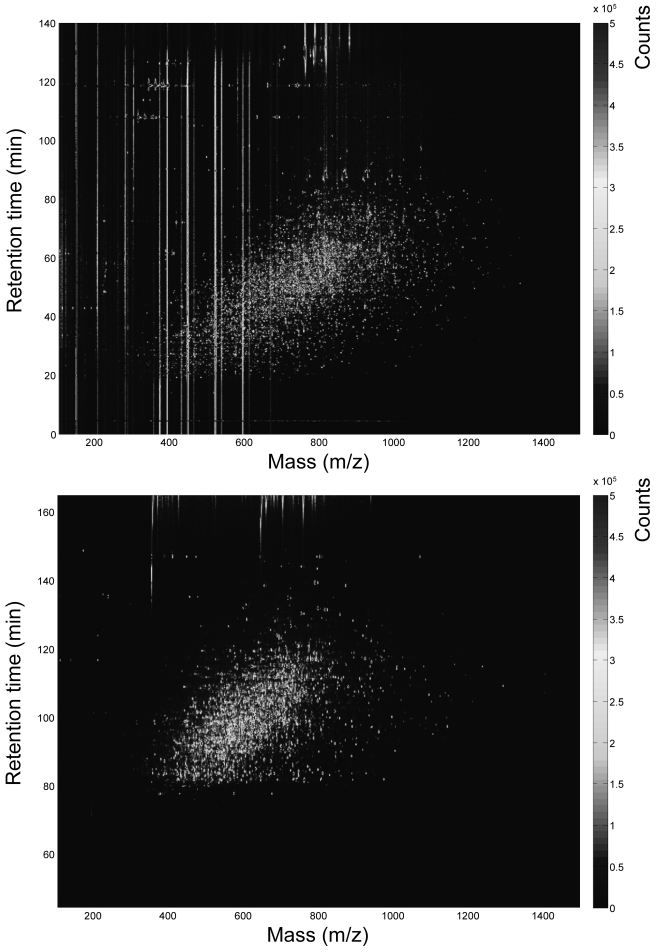

Figure 6. LC-MS images of trypsin-digested human serum samples depleted of the 6 most abundant proteins obtained with the Chip-LC-MS (electrospray) (a) or the Cap-LC-MS platform (ionspray) (b). See Horvatovich *et al.* [100] and the caption of Figure 5 for a description of the different LC-MS systems.

Attempts have been made in the field of proteomics with the framework program CORRA [235], which integrates the SuperHirn [83] and SpecArray [87] label-free quantitative data processing pipelines, and includes MS/MS identification based onSequest [151] with a range of R-based statistical tools. CORRA uses the Annotated Putative Peptide Markup Language (APML) format to integrate the different modules of the quantitative data processing pipelines with protein identification and statistical analysis. CORRA provides a user-friendly web interface and executes the different processing tasks on a local cluster. However CORRA cannot manage large amounts of diverse metadata, as needed for effective project management, and the integrated tools are limited to bioinformatics modules developed in a closely collaborating bioinformatics community. Developing further CORRA and APML or other similar initiatives should provide a breakthrough in using newly developed bioinformatics tools and therefore accelerate life science research.

4 REFERENCES

1. van der Greef, J.; Stroobant, P.; van der Heijden, R., The role of analytical sciences in medical systems biology. *Curr Opin Chem Biol* **2004,** 8, (5), 559-65.
2. Anderson, N. L.; Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **2002,** 1, (11), 845-67.
3. Schiess, R.; Wollscheid, B.; Aebersold, R., Targeted proteomic strategy for clinical biomarker discovery. *Mol Oncol* **2009,** 3, (1), 33-44.
4. Shen, Y.; Jacobs, J. M.; Camp, D. G., 2nd; Fang, R.; Moore, R. J.; Smith, R. D.; Xiao, W.; Davis, R. W.; Tompkins, R. G., Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal Chem* **2004,** 76, (4), 1134-44.
5. Pernemalm, M.; Lewensohn, R.; Lehtio, J., Affinity prefractionation for MS-based plasma proteomics. *Proteomics* **2009,** 9, (6), 1420-7.
6. Duncan, M. W.; Aebersold, R.; Caprioli, R. M., The pros and cons of peptide-centric proteomics. *Nat Biotechnol* **2010,** 28, (7), 659-64.
7. Duncan, M. W.; Yergey, A. L.; Patterson, S. D., Quantifying proteins by mass spectrometry: the selectivity of SRM is only part of the problem. *Proteomics* **2009,** 9, (5), 1124-7.
8. Horvatovich, P.; Govorukhina, N.; Bischoff, R., Biomarker discovery by proteomics: challenges not only for the analytical chemist. *Analyst* **2006,** 131, (11), 1193-6.
9. Horvatovich, P. L.; Bischoff, R., Current technological challenges in biomarker discovery and validation. *Eur J Mass Spectrom (Chichester, Eng)* **2010,** 16, (1), 101-21.
10. Guo, Y.; Fu, Z.; Van Eyk, J. E., A proteomic primer for the clinician. *Proc Am Thorac Soc* **2007,** 4, (1), 9-17.
11. Negishi, A.; Ono, M.; Handa, Y.; Kato, H.; Yamashita, K.; Honda, K.; Shitashige, M.; Satow, R.; Sakuma, T.; Kuwabara, H.; Omura, K.; Hirohashi, S.; Yamada, T., Large-scale quantitative clinical proteomics by label-free liquid chromatography and mass spectrometry. *Cancer Sci* **2009,** 100, (3), 514-9.
12. Balog, C. I.; Hensbergen, P. J.; Derks, R.; Verweij, J. J.; van Dam, G. J.; Vennervald, B. J.; Deelder, A. M.; Mayboroda, O. A., Novel automated biomarker discovery work flow for urinary peptidomics. *Clin Chem* **2009,** 55, (1), 117-25.
13. Ralhan, R.; Desouza, L. V.; Matta, A.; Chandra Tripathi, S.; Ghanny, S.; Datta Gupta, S.; Bahadur, S.; Siu, K. W., Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Mol Cell Proteomics* **2008,** 7, (6), 1162-73.
14. Higgs, R. E.; Knierman, M. D.; Gelfanova, V.; Butler, J. P.; Hale, J. E., Label-free LC-MS method for the identification of biomarkers. *Methods Mol Biol* **2008,** 428, 209-30.
15. Levin, Y.; Schwarz, E.; Wang, L.; Leweke, F. M.; Bahn, S., Label-free LC-MS/MS quantitative proteomics for large-scale biomarker discovery in complex samples. *J Sep Sci* **2007,** 30, (14), 2198-203.
16. Pan, S.; Zhang, H.; Rush, J.; Eng, J.; Zhang, N.; Patterson, D.; Comb, M. J.; Aebersold, R., High throughput proteome screening for biomarker detection. *Mol Cell Proteomics* **2005,** 4, (2), 182-90.
17. Gao, J.; Garulacan, L. A.; Storm, S. M.; Opiteck, G. J.; Dubaquie, Y.; Hefta, S. A.; Dambach, D. M.; Dongre, A. R., Biomarker discovery in biological fluids. *Methods* **2005,** 35, (3), 291-302.
18. Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A., Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **2004,** 3, (10), 984-97.
19. Bodovitz, S.; Joos, T., The proteomics bottleneck: strategies for preliminary validation of potential biomarkers and drug targets. *Trends Biotechnol* **2004,** 22, (1), 4-7.
20. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **2007,** 389, (4), 1017-31.

21. Gstaiger, M.; Aebersold, R., Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet* **2009,** 10, (9), 617-27.
22. Becker, G. W., Stable isotopic labeling of proteins for quantitative proteomic applications. *Brief Funct Genomic Proteomic* **2008,** 7, (5), 371-82.
23. Gevaert, K.; Impens, F.; Ghesquiere, B.; Van Damme, P.; Lambrechts, A.; Vandekerckhove, J., Stable isotopic labeling in proteomics. *Proteomics* **2008,** 8, (23-24), 4873-85.
24. Leitner, A.; Lindner, W., Chemistry meets proteomics: the use of chemical tagging reactions for MS-based proteomics. *Proteomics* **2006,** 6, (20), 5418-34.
25. Guerrera, I. C.; Kleiner, O., Application of mass spectrometry in proteomics. *Biosci Rep* **2005,** 25, (1-2), 71-93.
26. Yan, W.; Chen, S. S., Mass spectrometry-based quantitative proteomic profiling. *Brief Funct Genomic Proteomic* **2005,** 4, (1), 27-38.
27. Panchaud, A.; Affolter, M.; Moreillon, P.; Kussmann, M., Experimental and computational approaches to quantitative proteomics: status quo and outlook. *J Proteomics* **2008,** 71, (1), 19-33.
28. Buyse, M.; Sargent, D. J.; Grothey, A.; Matheson, A.; de Gramont, A., Biomarkers and surrogate end points--the challenge of statistical validation. *Nat Rev Clin Oncol* **2010,** 7, (6), 309-17.
29. Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R., 3rd; Bairoch, A.; Bergeron, J. J., Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods* **2010,** 7, (9), 681-5.
30. Scherf, U.; Becker, R.; Chan, M.; Hojvat, S., Approval of novel biomarkers: FDA's perspective and major requests. *Scand J Clin Lab Invest Suppl* **2010,** 242, 96-102.
31. Tan, D. S.; Thomas, G. V.; Garrett, M. D.; Banerji, U.; de Bono, J. S.; Kaye, S. B.; Workman, P., Biomarker-driven early clinical trials in oncology: a paradigm shift in drug development. *Cancer J* **2009,** 15, (5), 406-20.
32. Lin, D.; Hollander, Z.; Meredith, A.; McManus, B. M., Searching for 'omic' biomarkers. *Can J Cardiol* **2009,** 25 Suppl A, 9A-14A.
33. Dunn, B. K.; Wagner, P. D.; Anderson, D.; Greenwald, P., Molecular markers for early detection. *Semin Oncol* **2010,** 37, (3), 224-42.
34. Mischak, H.; Allmaier, G.; Apweiler, R.; Attwood, T.; Baumann, M.; Benigni, A.; Bennett, S. E.; Bischoff, R.; Bongcam-Rudloff, E.; Capasso, G.; Coon, J. J.; D'Haese, P.; Dominiczak, A. F.; Dakna, M.; Dihazi, H.; Ehrich, J. H.; Fernandez-Llama, P.; Fliser, D.; Frokiaer, J.; Garin, J.; Girolami, M.; Hancock, W. S.; Haubitz, M.; Hochstrasser, D.; Holman, R. R.; Ioannidis, J. P.; Jankowski, J.; Julian, B. A.; Klein, J. B.; Kolch, W.; Luider, T.; Massy, Z.; Mattes, W. B.; Molina, F.; Monsarrat, B.; Novak, J.; Peter, K.; Rossing, P.; Sanchez-Carbayo, M.; Schanstra, J. P.; Semmes, O. J.; Spasovski, G.; Theodorescu, D.; Thongboonkerd, V.; Vanholder, R.; Veenstra, T. D.; Weissinger, E.; Yamamoto, T.; Vlahou, A., Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* **2010,** 2, (46), 46ps42.
35. Chen, G.; Pramanik, B. N.; Liu, Y. H.; Mirza, U. A., Applications of LC/MS in structure identifications of small molecules and proteins in drug discovery. *J Mass Spectrom* **2007,** 42, (3), 279-87.
36. Chen, G.; Pramanik, B. N., Application of LC/MS to proteomics studies: current status and future prospects. *Drug Discov Today* **2009,** 14, (9-10), 465-71.
37. Rho, S.; You, S.; Kim, Y.; Hwang, D., From proteomics toward systems biology: integration of different types of proteomics data into network models. *BMB Rep* **2008,** 41, (3), 184-93.
38. Hoffmann, E. d., *Mass spectrometry : principles and applications*. 3rd ed. ed.; J. Wiley: Chichester, West Sussex, England ;, 2007.
39. Yates, J. R.; Ruse, C. I.; Nakorchevsky, A., Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* **2009,** 11, 49-79.
40. Ekman, R. S., J.; Westman-Brinkmalm, A.; Kraj, A., *Mass Spectrometry: Instrumentation, Interpretation, and Applications*. John Wiley & Sons, Inc: New Jersey, 2009.

41. Watson, J., Throck; Sparkman, O., David, *Introduction to Mass Spectrometry: Instrumentation, Applications and Strategies for Data Interpretation*. Fourth Edition ed.; John Wiley & Sons, Ltd.: Chichester, 2007; p 862.
42. Domon, B.; Aebersold, R., Mass spectrometry and protein analysis. *Science* **2006,** 312, (5771), 212-7.
43. Han, X.; Aslanian, A.; Yates, J. R., 3rd, Mass spectrometry for proteomics. *Curr Opin Chem Biol* **2008,** 12, (5), 483-90.
44. Domon, B.; Aebersold, R., Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* **2010,** 28, (7), 710-21.
45. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** 76, (14), 4193-201.
46. Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G., Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **2005,** 4, (10), 1487-502.
47. Grossmann, J.; Roschitzki, B.; Panse, C.; Fortes, C.; Barkow-Oesterreicher, S.; Rutishauser, D.; Schlapbach, R., Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J Proteomics* **2010,** 73, (9), 1740-6.
48. Mueller, L. N.; Brusniak, M. Y.; Mani, D. R.; Aebersold, R., An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res* **2008,** 7, (1), 51-61.
49. Li, M.; Gray, W.; Zhang, H.; Chung, C. H.; Billheimer, D.; Yarbrough, W. G.; Liebler, D. C.; Shyr, Y.; Slebos, R. J., Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J Proteome Res* **2010,** 9, (8), 4295-305.
50. Zhu, W.; Smith, J. W.; Huang, C. M., Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol* 2010, 840518.
51. Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **2005,** 4, (9), 1265-72.
52. Shinoda, K.; Tomita, M.; Ishihama, Y., emPAI Calc--for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry. *Bioinformatics* **2010,** 26, (4), 576-7.
53. Braisted, J. C.; Kuntumalla, S.; Vogel, C.; Marcotte, E. M.; Rodrigues, A. R.; Wang, R.; Huang, S. T.; Ferlanti, E. S.; Saeed, A. I.; Fleischmann, R. D.; Peterson, S. N.; Pieper, R., The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* **2008,** 9, 529.
54. Lu, P.; Vogel, C.; Wang, R.; Yao, X.; Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **2007,** 25, (1), 117-24.
55. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **2006,** 5, (1), 144-56.
56. Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A. J.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C., Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res* **2010,** 9, (2), 761-76.
57. Schmidt, A.; Claassen, M.; Aebersold, R., Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr Opin Chem Biol* **2009,** 13, (5-6), 510-7.
58. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009,** 138, (4), 795-806.

59. Picotti, P.; Rinner, O.; Stallmach, R.; Dautel, F.; Farrah, T.; Domon, B.; Wenschuh, H.; Aebersold, R., High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat Methods* **2010,** 7, (1), 43-6.
60. Pan, S.; Aebersold, R.; Chen, R.; Rush, J.; Goodlett, D. R.; McIntosh, M. W.; Zhang, J.; Brentnall, T. A., Mass spectrometry based targeted protein quantification: methods and applications. *J Proteome Res* **2009,** 8, (2), 787-97.
61. Deutsch, E. W.; Lam, H.; Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **2008,** 9, (5), 429-34.
62. Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R., Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **2007,** 25, (1), 125-31.
63. Sherwood, C. A.; Eastham, A.; Lee, L. W.; Peterson, A.; Eng, J. K.; Shteynberg, D.; Mendoza, L.; Deutsch, E. W.; Risler, J.; Tasman, N.; Aebersold, R.; Lam, H.; Martin, D. B., MaRiMba: a software application for spectral library-based MRM transition list assembly. *J Proteome Res* **2009,** 8, (10), 4396-405.
64. Yang, X.; Lazar, I. M., MRM screening/biomarker discovery with linear ion trap MS: a library of human cancer-specific peptides. *BMC Cancer* **2009,** 9, 96.
65. Huttenhain, R.; Malmstrom, J.; Picotti, P.; Aebersold, R., Perspectives of targeted mass spectrometry for protein biomarker verification. *Curr Opin Chem Biol* **2009,** 13, (5-6), 518-25.
66. Lange, V.; Picotti, P.; Domon, B.; Aebersold, R., Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **2008,** 4, 222.
67. Orchard, S.; Jones, P.; Taylor, C.; Zhu, W.; Julian, R. K., Jr.; Hermjakob, H.; Apweiler, R., Proteomic data exchange and storage: the need for common standards and public repositories. *Methods Mol Biol* **2007,** 367, 261-70.
68. Orchard, S.; Taylor, C.; Hermjakob, H.; Zhu, W.; Julian, R.; Apweiler, R., Current status of proteomic standards development. *Expert Rev Proteomics* **2004,** 1, (2), 179-83.
69. Cote, R. G.; Reisinger, F.; Martens, L., jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* **2010,** 10, (7), 1332-5.
70. Deutsch, E., mzML: a single, unifying data format for mass spectrometer output. *Proteomics* **2008,** 8, (14), 2776-7.
71. Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Rompp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P. A.; Deutsch, E. W., mzML - a Community Standard for Mass Spectrometry Data. *Mol Cell Proteomics* **2010**.
72. Taylor, C. F., Minimum reporting requirements for proteomics: a MIAPE primer. *Proteomics* **2006,** 6 Suppl 2, 39-44.
73. Taylor, C. F., Standards for reporting bioscience data: a forward look. *Drug Discov Today* **2007,** 12, (13-14), 527-33.
74. Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P. A.; Julian, R. K., Jr.; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; Dunn, M. J.; Heck, A. J.; Leitner, A.; Macht, M.; Mann, M.; Martens, L.; Neubert, T. A.; Patterson, S. D.; Ping, P.; Seymour, S. L.; Souda, P.; Tsugita, A.; Vandekerckhove, J.; Vondriska, T. M.; Whitelegge, J. P.; Wilkins, M. R.; Xenarios, I.; Yates, J. R., 3rd; Hermjakob, H., The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* **2007,** 25, (8), 887-93.
75. Monroe, M. E.; Tolic, N.; Jaitly, N.; Shaw, J. L.; Adkins, J. N.; Smith, R. D., VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* **2007,** 23, (15), 2021-3.
76. Kohlbacher, O.; Reinert, K.; Gropl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M., TOPP--the OpenMS proteomics pipeline. *Bioinformatics* **2007,** 23, (2), e191-7.
77. Lange, E.; Gropl, C.; Schulz-Trieglaff, O.; Leinenbach, A.; Huber, C.; Reinert, K., A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics* **2007,** 23, (13), i273-81.
78. Reinert, K.; Kohlbacher, O., OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **2010,** 604, 201-11.

79. Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O., OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008,** 9, 163.
80. Katajamaa, M.; Miettinen, J.; Oresic, M., MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006,** 22, (5), 634-6.
81. Katajamaa, M.; Oresic, M., Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* **2005,** 6, 179.
82. Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R., Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* **2001,** 19, (10), 946-51.
83. Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M. Y.; Vitek, O.; Aebersold, R.; Muller, M., SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **2007,** 7, (19), 3470-80.
84. Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R., 3rd, A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods* **2008,** 5, (4), 319-22.
85. Leptos, K. C.; Sarracino, D. A.; Jaffe, J. D.; Krastins, B.; Church, G. M., MapQuant: open-source software for large-scale protein quantification. *Proteomics* **2006,** 6, (6), 1770-82.
86. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech* **2008,** 26, (12), 1367-1372.
87. Li, X. J.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R., A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteomics* **2005,** 4, (9), 1328-40.
88. Meiring, H. D.; Soethout, E. C.; de Jong, A. P.; van Els, C. A., Targeted identification of infection-related HLA class I-presented epitopes by stable isotope tagging of epitopes (SITE). *Curr Protoc Immunol* **2007,** 77, 16.3.1-16.3.20.
89. Jaffe, J. D.; Mani, D. R.; Leptos, K. C.; Church, G. M.; Gillette, M. A.; Carr, S. A., PEPPeR, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics* **2006,** 5, (10), 1927-41.
90. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G., XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* **2006,** 78, (3), 779-87.
91. Tautenhahn, R.; Bottcher, C.; Neumann, S., Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **2008,** 9, 504.
92. Aberg, K. M.; Torgrip, R. J.; Kolmert, J.; Schuppe-Koistinen, I.; Lindberg, J., Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of pure ion chromatograms using Kalman tracking. *J Chromatogr A* **2008,** 1192, (1), 139-46.
93. Stolt, R.; Torgrip, R. J.; Lindberg, J.; Csenki, L.; Kolmert, J.; Schuppe-Koistinen, I.; Jacobsson, S. P., Second-order peak detection for multicomponent high-resolution LC/MS data. *Anal Chem* **2006,** 78, (4), 975-83.
94. Windig, W.; Phalp, J. M.; Payne, A. W., A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry* **1996,** 68, (20), 3602-3606.
95. Christin, C.; Hoefsloot, H. C.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L., Time Alignment Algorithms Based on Selected Mass Traces for Complex LC-MS Data. *J Proteome Res* **2010**.
96. Christin, C.; Smilde, A. K.; Hoefsloot, H. C.; Suits, F.; Bischoff, R.; Horvatovich, P. L., Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal Chem* **2008,** 80, (18), 7012-21.
97. Du, P.; Stolovitzky, G.; Horvatovich, P.; Bischoff, R.; Lim, J.; Suits, F., A noise model for mass spectrometry based proteomics. *Bioinformatics* **2008,** 24, (8), 1070-7.
98. Busch, K. L., Chemical Noise in Mass Spectrometry. *Spectroscopy* **2002,** 17, (10), 6.
99. Annesley, T. M., Ion suppression in mass spectrometry. *Clin Chem* **2003,** 49, (7), 1041-4.

100. Horvatovich, P.; Govorukhina, N. I.; Reijmers, T. H.; van der Zee, A. G.; Suits, F.; Bischoff, R., Chip-LC-MS for label-free profiling of human serum. *Electrophoresis* **2007,** 28, (23), 4493-505.
101. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **2003,** 75, (18), 4818-26.
102. Li, Y.; Qu, H.; Cheng, Y., An entropy-based method for noise reduction of liquid chromatography-mass spectrometry data. *Anal Chim Acta* **2008,** 612, (1), 19-22.
103. Listgarten, J.; Emili, A., Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **2005,** 4, (4), 419-34.
104. Yu, T.; Park, Y.; Johnson, J. M.; Jones, D. P., apLCMS--adaptive processing of high-resolution LC/MS data. *Bioinformatics* **2009,** 25, (15), 1930-6.
105. Andreev, V. P.; Rejtar, T.; Chen, H. S.; Moskovets, E. V.; Ivanov, A. R.; Karger, B. L., A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal Chem* **2003,** 75, (22), 6314-26.
106. Cappadona, S.; Levander, F.; Jansson, M.; James, P.; Cerutti, S.; Pattini, L., Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry. *Anal Chem* **2008,** 80, (13), 4960-8.
107. Cappadona, S.; Nanni, P.; Benevento, M.; Levander, F.; Versura, P.; Roda, A.; Cerutti, S.; Pattini, L., Improved label-free LC-MS analysis by wavelet-based noise rejection. *J Biomed Biotechnol* **2010,** 2010, 131505.
108. Du, P.; Sudha, R.; Prystowsky, M. B.; Angeletti, R. H., Data reduction of isotope-resolved LC-MS spectra. *Bioinformatics* **2007,** 23, (11), 1394-400.
109. Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M., A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **2006,** 22, (15), 1902-9.
110. Zhang, J.; Gonzalez, E.; Hestilow, T.; Haskins, W.; Huang, Y., Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr Genomics* **2009,** 10, (6), 388-401.
111. de Groot, J. C.; Fiers, M. W.; van Ham, R. C.; America, A. H., Post alignment clustering procedure for comparative quantitative proteomics LC-MS data. *Proteomics* **2008,** 8, (1), 32-6.
112. Piening, B. D.; Wang, P.; Bangur, C. S.; Whiteaker, J.; Zhang, H.; Feng, L. C.; Keane, J. F.; Eng, J. K.; Tang, H.; Prakash, A.; McIntosh, M. W.; Paulovich, A., Quality control metrics for LC-MS feature detection tools demonstrated on Saccharomyces cerevisiae proteomic profiles. *J Proteome Res* **2006,** 5, (7), 1527-34.
113. Zimmer, J. S.; Monroe, M. E.; Qian, W. J.; Smith, R. D., Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev* **2006,** 25, (3), 450-82.
114. Scheltema, R. A.; Kamleh, A.; Wildridge, D.; Ebikeme, C.; Watson, D. G.; Barrett, M. P.; Jansen, R. C.; Breitling, R., Increasing the mass accuracy of high-resolution LC-MS data using background ions: a case study on the LTQ-Orbitrap. *Proteomics* **2008,** 8, (22), 4647-56.
115. Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P., Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol Cell Proteomics* **2006,** 5, (7), 1326-37.
116. Danell, R. M.; Ouvry-Patat, S. A.; Scarlett, C. O.; Speir, J. P.; Borchers, C. H., Data Self-Recalibration and Mixture Mass Fingerprint Searching (DASER-MMF) to enhance protein identification within complex mixtures. *J Am Soc Mass Spectrom* **2008,** 19, (12), 1914-25.
117. Dijkstra, M.; Jansen, R. C., Optimal analysis of complex protein mass spectra. *Proteomics* **2009,** 9, (15), 3869-76.
118. Kemperman, R. F.; Horvatovich, P. L.; Hoekman, B.; Reijmers, T. H.; Muskiet, F. A.; Bischoff, R., Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: method development, evaluation, and application to proteinuria. *J Proteome Res* **2007,** 6, (1), 194-206.

119. Yang, Y. H.; Dudoit, S.; Luu, P.; Lin, D. M.; Peng, V.; Ngai, J.; Speed, T. P., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **2002,** 30, (4), e15.
120. Kultima, K.; Nilsson, A.; Scholz, B.; Rossbach, U. L.; Falth, M.; Andren, P. E., Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol Cell Proteomics* **2009,** 8, (10), 2285-95.
121. Callister, S. J.; Barry, R. C.; Adkins, J. N.; Johnson, E. T.; Qian, W. J.; Webb-Robertson, B. J.; Smith, R. D.; Lipton, M. S., Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* **2006,** 5, (2), 277-86.
122. Karpievitch, Y. V.; Taverner, T.; Adkins, J. N.; Callister, S. J.; Anderson, G. A.; Smith, R. D.; Dabney, A. R., Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics* **2009,** 25, (19), 2573-80.
123. Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J Chromatogr A* **2002,** 961, (2), 237-44.
124. Clifford, D.; Stone, G.; Montoliu, I.; Rezzi, S.; Martin, F. o.-P.; Guy, P.; Bruce, S.; Kochhar, S., Alignment Using Variable Penalty Dynamic Time Warping. *Anal Chem* **2009**.
125. Eilers, P. H., Parametric time warping. *Anal Chem* **2004,** 76, (2), 404-11.
126. Kassidas, A.; MacGregor, J. F.; Taylor, P. A., Synchronization of batch trajectories using dynamic time warping. *AIChe* **1998,** 44, 864.
127. Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. r., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J Chromatogr A* **1998,** 805, 17-35.
128. Ramaker, H.-J.; van Sprang, E. N. M.; Westerhuis, J. A.; Smilde, A. K., Dynamic time warping of spectroscopic BATCH data. *Anal Chim Acta* **2003,** 498, 133-153.
129. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemometrics* **2004,** 18, 231-241.
130. van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H.; Heyden, Y. V., A comparison of three algorithms for chromatograms alignment. *J Chromatogr A* **2006,** 1118, (2), 199-210.
131. Vial, J.; Nocairi, H.; Sassiat, P.; Mallipatu, S.; Cognon, G.; Thiebaut, D.; Teillet, B.; Rutledge, D. N., Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. *J Chromatogr A* **2009,** 1216, (14), 2866-72.
132. Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B., Signal maps for mass spectrometry-based comparative proteomics. *Mol Cell Proteomics* **2006,** 5, (3), 423-32.
133. Prince, J. T.; Marcotte, E. M., Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal Chem* **2006,** 78, (17), 6140-52.
134. Sadygov, R. G.; Maroto, F. M.; Huhmer, A. F., ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Anal Chem* **2006,** 78, (24), 8207-17.
135. Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stuhler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenfuhrer, J., Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **2009,** 25, (6), 758-64.
136. Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P., Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal Chem* **2008,** 80, (9), 3095-104.
137. Jaitly, N.; Monroe, M. E.; Petyuk, V. A.; Clauss, T. R.; Adkins, J. N.; Smith, R. D., Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal Chem* **2006,** 78, (21), 7397-409.
138. Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C., Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **2008,** 9, 375.

139. Listgarten, J.; Neal, R. M.; Roweis, S. T.; Wong, P.; Emili, A., Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics* **2007,** 23, (2), e198-204.
140. Tsou, C. C.; Tsai, C. F.; Tsui, Y. H.; Sudhir, P. R.; Wang, Y. T.; Chen, Y. J.; Chen, J. Y.; Sung, T. Y.; Hsu, W. L., IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Mol Cell Proteomics* **2010,** 9, (1), 131-44.
141. Aberg, K. M.; Alm, E.; Torgrip, R. J., The correspondence problem for metabonomics datasets. *Anal Bioanal Chem* **2009,** 394, (1), 151-62.
142. De Beer, M.; Lynen, F.; Chen, K.; Ferguson, P.; Hanna-Brown, M.; Sandra, P., Stationary-phase optimized selectivity liquid chromatography: development of a linear gradient prediction algorithm. *Anal Chem* **2010,** 82, (5), 1733-43.
143. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J Sep Sci* **2005,** 28, (14), 1694-703.
144. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem* **2005,** 77, (19), 6426-34.
145. Vandenbogaert, M.; Li-Thiao-Te, S.; Kaltenbach, H. M.; Zhang, R.; Aittokallio, T.; Schwikowski, B., Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics* **2008,** 8, (4), 650-72.
146. Prakash, A.; Piening, B.; Whiteaker, J.; Zhang, H.; Shaffer, S. A.; Martin, D.; Hohmann, L.; Cooke, K.; Olson, J. M.; Hansen, S.; Flory, M. R.; Lee, H.; Watts, J.; Goodlett, D. R.; Aebersold, R.; Paulovich, A.; Schwikowski, B., Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. *Mol Cell Proteomics* **2007,** 6, (10), 1741-8.
147. Hilario, M.; Kalousis, A.; Pellegrini, C.; Muller, M., Processing and classification of protein mass spectra. *Mass Spectrom Rev* **2006,** 25, (3), 409-49.
148. Kapp, E.; Schutz, F., Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Curr Protoc Protein Sci* **2007,** 49, 25.2.1-25.2.19.
149. Nesvizhskii, A. I., Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol* **2007,** 367, 87-119.
150. Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd, Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* **2004,** 1, (3), 195-202.
151. Eng, J. K.; McCormack, A. L.; Yates Iii, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** 5, (11), 976-989.
152. Mann, M.; Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* **1994,** 66, (24), 4390-9.
153. Clauser, K. R.; Baker, P.; Burlingame, A. L., Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* **1999,** 71, (14), 2871-82.
154. Tabb, D. L.; Saraf, A.; Yates, J. R., 3rd, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* **2003,** 75, (23), 6415-21.
155. Bafna, V.; Edwards, N., SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **2001,** 17 Suppl 1, S13-21.
156. Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J., OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003,** 3, (8), 1454-63.
157. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** 20, (18), 3551-67.
158. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *J Proteome Res* **2004,** 3, (5), 958-64.
159. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R., Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007,** 7, (5), 655-67.

160. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R., Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods* **2008,** 5, (10), 873-5.
161. Frank, A. M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S. P.; Smith, R. D.; Pevzner, P. A., Clustering millions of tandem mass spectra. *J Proteome Res* **2008,** 7, (1), 113-22.
162. Hernandez, P.; Gras, R.; Frey, J.; Appel, R. D., Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* **2003,** 3, (6), 870-8.
163. Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V., InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **2005,** 77, (14), 4626-39.
164. Ma, B.; Lajoie, G., De novo interpretation of tandem mass spectra. *Curr Protoc Bioinformatics* **2009,** 25, 13.10.1-13.10.8.
165. Tannu, N. S.; Hemby, S. E., De novo protein sequence analysis of Macaca mulatta. *BMC Genomics* **2007,** 8, 270.
166. Frank, A.; Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **2005,** 77, (4), 964-73.
167. Bern, M.; Goldberg, D., De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J Comput Biol* **2006,** 13, (2), 364-78.
168. Taylor, J. A.; Johnson, R. S., Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* **2001,** 73, (11), 2594-604.
169. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A., De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* **1999,** 6, (3-4), 327-42.
170. Mo, L.; Dutta, D.; Wan, Y.; Chen, T., MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal Chem* **2007,** 79, (13), 4870-8.
171. DiMaggio, P. A., Jr.; Floudas, C. A., De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal Chem* **2007,** 79, (4), 1433-46.
172. Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M., NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem* **2005,** 77, (22), 7265-73.
173. Grossmann, J.; Roos, F. F.; Cieliebak, M.; Liptak, Z.; Mathis, L. K.; Muller, M.; Gruissem, W.; Baginsky, S., AUDENS: a tool for automated peptide de novo sequencing. *J Proteome Res* **2005,** 4, (5), 1768-74.
174. Norbeck, A. D.; Monroe, M. E.; Adkins, J. N.; Anderson, K. K.; Daly, D. S.; Smith, R. D., The utility of accurate mass and LC elution time information in the analysis of complex proteomes. *J Am Soc Mass Spectrom* **2005,** 16, (8), 1239-49.
175. Masselon, C. D.; Kieffer-Jaquinod, S.; Brugiere, S.; Dupierris, V.; Garin, J., Influence of mass resolution on species matching in accurate mass and retention time (AMT) tag proteomics experiments. *Rapid Commun Mass Spectrom* **2008,** 22, (7), 986-92.
176. Kim, Y. J.; Feild, B.; Fitzhugh, W.; Heidbrink, J. L.; Duff, J. W.; Heil, J.; Ruben, S. M.; He, T., Reference map for liquid chromatography-mass spectrometry-based quantitative proteomics. *Anal Biochem* **2009,** 393, (2), 155-62.
177. Searle, B. C., Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **2010,** 10, (6), 1265-1269.
178. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.
179. Bjornson, R. D.; Carriero, N. J.; Colangelo, C.; Shifman, M.; Cheung, K. H.; Miller, P. L.; Williams, K., X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *J Proteome Res* **2008,** 7, (1), 293-9.
180. Baczek, T.; Kaliszan, R., Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. *Proteomics* **2009,** 9, (4), 835-47.
181. Gilar, M.; Jaworski, A.; Olivova, P.; Gebler, J. C., Peptide retention prediction applied to proteomic data analysis. *Rapid Commun Mass Spectrom* **2007,** 21, (17), 2813-21.

182. Shinoda, K.; Sugimoto, M.; Tomita, M.; Ishihama, Y., Informatics for peptide retention properties in proteomic LC-MS. *Proteomics* **2008,** 8, (4), 787-98.
183. Quandt, A.; Masselot, A.; Hernandez, P.; Hernandez, C.; Maffioletti, S.; Appel, R. D.; Lisacek, F., SwissPIT: An workflow-based platform for analyzing tandem-MS spectra using the Grid. *Proteomics* **2009,** 9, (10), 2648-55.
184. Quandt, A.; Hernandez, P.; Masselot, A.; Hernandez, C.; Maffioletti, S.; Pautasso, C.; Appel, R. D.; Lisacek, F., swissPIT: a novel approach for pipelined analysis of mass spectrometry data. *Bioinformatics* **2008,** 24, (11), 1416-7.
185. Bandeira, N.; Tsur, D.; Frank, A.; Pevzner, P. A., Protein identification by spectral networks analysis. *Proc Natl Acad Sci U S A* **2007,** 104, (15), 6140-5.
186. Bandeira, N.; Olsen, J. V.; Mann, J. V.; Mann, M.; Pevzner, P. A., Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics* **2008,** 24, (13), i416-23.
187. Kim, S.; Gupta, N.; Bandeira, N.; Pevzner, P. A., Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol Cell Proteomics* **2009,** 8, (1), 53-69.
188. Bianco, L.; Mead, J.; Bessant, C., Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS datasets. *J Proteome Res* **2009**.
189. Fitzgibbon, M.; Li, Q.; McIntosh, M., Modes of inference for evaluating the confidence of peptide identifications. *J Proteome Res* **2008,** 7, (1), 35-9.
190. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* **2010,** 604, 55-71.
191. Pendarvis, K.; Kumar, R.; Burgess, S. C.; Nanduri, B., An automated proteomic data analysis workflow for mass spectrometry. *BMC Bioinformatics* **2009,** 10 Suppl 11, S17.
192. Hughes, C.; Ma, B.; Lajoie, G. A., De novo sequencing methods in proteomics. *Methods Mol Biol* **2010,** 604, 105-21.
193. Jones, S. J. H. A. R., *Proteome Bioinformatics*. Humana Press: Hatfield, 2010; Vol. 604.
194. Xu, C.; Ma, B., Software for computational peptide identification from MS-MS data. *Drug Discov Today* **2006,** 11, (13-14), 595-600.
195. Salmi, J.; Nyman, T. A.; Nevalainen, O. S.; Aittokallio, T., Filtering strategies for improving protein identification in high-throughput MS/MS studies. *Proteomics* **2009,** 9, (4), 848-60.
196. Barton, S. J.; Whittaker, J. C., Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrom Rev* **2009,** 28, (1), 177-87.
197. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L., Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal Chem* **2010,** 82, (6), 2272-81.
198. Podwojski, K.; Eisenacher, M.; Kohl, M.; Turewicz, M.; Meyer, H. E.; Rahnenfuhrer, J.; Stephan, C., Peek a peak: a glance at statistics for quantitative label-free proteomics. *Expert Rev Proteomics* **2010,** 7, (2), 249-61.
199. Lau, K. W.; Jones, A. R.; Swainston, N.; Siepen, J. A.; Hubbard, S. J., Capture and analysis of quantitative proteomic data. *Proteomics* **2007,** 7, (16), 2787-99.
200. Kim, H.; Eliuk, S.; Deshane, J.; Meleth, S.; Sanderson, T.; Pinner, A.; Robinson, G.; Wilson, L.; Kirk, M.; Barnes, S., 2D gel proteomics: an approach to study age-related differences in protein abundance or isoform complexity in biological samples. *Methods Mol Biol* **2007,** 371, 349-91.
201. Lopez, J. L., Two-dimensional electrophoresis in proteome expression analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* **2007,** 849, (1-2), 190-202.
202. Marengo, E.; Robotti, E.; Bobba, M., 2D-PAGE maps analysis. *Methods Mol Biol* **2008,** 428, 291-325.
203. Lv, L. L.; Liu, B. C., High-throughput antibody microarrays for quantitative proteomic analysis. *Expert Rev Proteomics* **2007,** 4, (4), 505-13.

204. Reid, J. D.; Parker, C. E.; Borchers, C. H., Protein arrays for biomarker discovery. *Curr Opin Mol Ther* **2007,** 9, (3), 216-21.
205. VanMeter, A.; Signore, M.; Pierobon, M.; Espina, V.; Liotta, L. A.; Petricoin, E. F., 3rd, Reverse-phase protein microarrays: application to biomarker discovery and translational medicine. *Expert Rev Mol Diagn* **2007,** 7, (5), 625-33.
206. Basak, S. C.; Gute, B. D., Mathematical biodescriptors of proteomics maps: background and applications. *Curr Opin Drug Discov Devel* **2008,** 11, (3), 320-6.
207. Eriksson, L.; Johansson, E.; Lindgren, F.; Sjostrom, M.; Wold, S., Megavariate analysis of hierarchical QSAR data. *J Comput Aided Mol Des* **2002,** 16, (10), 711-26.
208. Ronald, E. S., Multi- and Megavariate Data Analysis. Principles and Applications, I. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Umetrics Academy, Umeå, 2001, ISBN 91-973730-1-X, 533pp. *Journal of Chemometrics* **2002,** 16, (5), 261-262.
209. Byvatov, E.; Schneider, G., Support vector machine applications in bioinformatics. *Appl Bioinformatics* **2003,** 2, (2), 67-77.
210. Li, L.; Tang, H.; Wu, Z.; Gong, J.; Gruidl, M.; Zou, J.; Tockman, M.; Clark, R. A., Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med* **2004,** 32, (2), 71-83.
211. Rajapakse, J. C.; Duan, K. B.; Yeo, W. K., Proteomic cancer classification with mass spectrometry data. *Am J Pharmacogenomics* **2005,** 5, (5), 281-92.
212. Torrecilla, J. S.; Rojo, E.; Oliet, M.; Dominguez, J. C.; Rodriguez, F., Self-organizing maps and learning vector quantization networks as tools to identify vegetable oils. *J Agric Food Chem* **2009,** 57, (7), 2763-9.
213. Villmann, T.; Schleif, F. M.; Kostrzewa, M.; Walch, A.; Hammer, B., Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Brief Bioinform* **2008,** 9, (2), 129-43.
214. Issaq, H. J.; Van, Q. N.; Waybright, T. J.; Muschik, G. M.; Veenstra, T. D., Analytical and statistical approaches to metabolomics research. *J Sep Sci* **2009,** 32, (13), 2183-99.
215. Smit, S.; Hoefsloot, H. C.; Smilde, A. K., Statistical data processing in clinical proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **2008,** 866, (1-2), 77-88.
216. Hendriks, M. M.; Smit, S.; Akkermans, W. L.; Reijmers, T. H.; Eilers, P. H.; Hoefsloot, H. C.; Rubingh, C. M.; de Koster, C. G.; Aerts, J. M.; Smilde, A. K., How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. *Proteomics* **2007,** 7, (20), 3672-80.
217. Smit, S.; van Breemen, M. J.; Hoefsloot, H. C.; Smilde, A. K.; Aerts, J. M.; de Koster, C. G., Assessing the statistical validity of proteomics based biomarkers. *Anal Chim Acta* **2007,** 592, (2), 210-7.
218. Reunanen, J., Search Strategies. In *Feature Extraction*, 2006; pp 119-136.
219. P. Silcocks, X. H. Z., N. Obuchowski N, D. McClish;, *Statistical methods in diagnostic medicine*. Wiley & Sons Interscience: New York, 2002.
220. Stead, D. A.; Paton, N. W.; Missier, P.; Embury, S. M.; Hedeler, C.; Jin, B.; Brown, A. J.; Preece, A., Information quality in proteomics. *Brief Bioinform* **2008,** 9, (2), 174-88.
221. Roelofsen, H.; Alvarez-Llamas, G.; Dijkstra, M.; Breitling, R.; Havenga, K.; Bijzet, J.; Zandbergen, W.; de Vries, M. P.; Ploeg, R. J.; Vonk, R. J., Analyses of intricate kinetics of the serum proteome during and after colon surgery by protein expression time series. *Proteomics* **2007,** 7, (17), 3219-28.
222. Vis, D. J.; Westerhuis, J. A.; Smilde, A. K.; van der Greef, J., Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics* **2007,** 8, 322.
223. Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C.; Lamers, R. J.; van der Greef, J.; Timmerman, M. E., ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **2005,** 21, (13), 3043-8.
224. Nueda, M. J.; Conesa, A.; Westerhuis, J. A.; Hoefsloot, H. C.; Smilde, A. K.; Talon, M.; Ferrer, A., Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics* **2007,** 23, (14), 1792-800.

225. Zhang, R.; Barton, A.; Brittenden, J.; Huang, J. T. J.; Crowther, D., Evaluation for computational platforms of LC-MS based label-free quantitative proteomics: global view. *Journal of Proteomics and Bioinformatics* **2010,** 3, (9), 6.

226. Paulovich, A. G.; Billheimer, D.; Ham, A. J.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C., Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics* **2010,** 9, (2), 242-54.

227. Kandasamy, K.; Keerthikumar, S.; Goel, R.; Mathivanan, S.; Patankar, N.; Shafreen, B.; Renuse, S.; Pawar, H.; Ramachandra, Y. L.; Acharya, P. K.; Ranganathan, P.; Chaerkady, R.; Keshava Prasad, T. S.; Pandey, A., Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res* **2009,** 37, (Database issue), D773-81.

228. Mathivanan, S.; Ahmed, M.; Ahn, N. G.; Alexandre, H.; Amanchy, R.; Andrews, P. C.; Bader, J. S.; Balgley, B. M.; Bantscheff, M.; Bennett, K. L.; Bjorling, E.; Blagoev, B.; Bose, R.; Brahmachari, S. K.; Burlingame, A. S.; Bustelo, X. R.; Cagney, G.; Cantin, G. T.; Cardasis, H. L.; Celis, J. E.; Chaerkady, R.; Chu, F.; Cole, P. A.; Costello, C. E.; Cotter, R. J.; Crockett, D.; DeLany, J. P.; De Marzo, A. M.; DeSouza, L. V.; Deutsch, E. W.; Dransfield, E.; Drewes, G.; Droit, A.; Dunn, M. J.; Elenitoba-Johnson, K.; Ewing, R. M.; Van Eyk, J.; Faca, V.; Falkner, J.; Fang, X.; Fenselau, C.; Figeys, D.; Gagne, P.; Gelfi, C.; Gevaert, K.; Gimble, J. M.; Gnad, F.; Goel, R.; Gromov, P.; Hanash, S. M.; Hancock, W. S.; Harsha, H. C.; Hart, G.; Hays, F.; He, F.; Hebbar, P.; Helsens, K.; Hermeking, H.; Hide, W.; Hjerno, K.; Hochstrasser, D. F.; Hofmann, O.; Horn, D. M.; Hruban, R. H.; Ibarrola, N.; James, P.; Jensen, O. N.; Jensen, P. H.; Jung, P.; Kandasamy, K.; Kheterpal, I.; Kikuno, R. F.; Korf, U.; Korner, R.; Kuster, B.; Kwon, M. S.; Lee, H. J.; Lee, Y. J.; Lefevre, M.; Lehvaslaiho, M.; Lescuyer, P.; Levander, F.; Lim, M. S.; Lobke, C.; Loo, J. A.; Mann, M.; Martens, L.; Martinez-Heredia, J.; McComb, M.; McRedmond, J.; Mehrle, A.; Menon, R.; Miller, C. A.; Mischak, H.; Mohan, S. S.; Mohmood, R.; Molina, H.; Moran, M. F.; Morgan, J. D.; Moritz, R.; Morzel, M.; Muddiman, D. C.; Nalli, A.; Navarro, J. D.; Neubert, T. A.; Ohara, O.; Oliva, R.; Omenn, G. S.; Oyama, M.; Paik, Y. K.; Pennington, K.; Pepperkok, R.; Periaswamy, B.; Petricoin, E. F.; Poirier, G. G.; Prasad, T. S.; Purvine, S. O.; Rahiman, B. A.; Ramachandran, P.; Ramachandra, Y. L.; Rice, R. H.; Rick, J.; Ronnholm, R. H.; Salonen, J.; Sanchez, J. C.; Sayd, T.; Seshi, B.; Shankari, K.; Sheng, S. J.; Shetty, V.; Shivakumar, K.; Simpson, R. J.; Sirdeshmukh, R.; Siu, K. W.; Smith, J. C.; Smith, R. D.; States, D. J.; Sugano, S.; Sullivan, M.; Superti-Furga, G.; Takatalo, M.; Thongboonkerd, V.; Trinidad, J. C.; Uhlen, M.; Vandekerckhove, J.; Vasilescu, J.; Veenstra, T. D.; Vidal-Taboada, J. M.; Vihinen, M.; Wait, R.; Wang, X.; Wiemann, S.; Wu, B.; Xu, T.; Yates, J. R.; Zhong, J.; Zhou, M.; Zhu, Y.; Zurbig, P.; Pandey, A., Human Proteinpedia enables sharing of human protein data. *Nat Biotechnol* **2008,** 26, (2), 164-7.

229. Mathivanan, S.; Pandey, A., Human Proteinpedia as a resource for clinical proteomics. *Mol Cell Proteomics* **2008,** 7, (10), 2038-47.

230. Peters, S.; van Velzen, E.; Janssen, H. G., Parameter selection for peak alignment in chromatographic sample profiling: objective quality indicators and use of control samples. *Anal Bioanal Chem* **2009,** 394, (5), 1273-81.

231. Taylor, J.; Schenck, I.; Blankenberg, D.; Nekrutenko, A., Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics* **2007,** 19, 10.5.1-10.5.25.

232. Blankenberg, D.; Taylor, J.; Schenck, I.; He, J.; Zhang, Y.; Ghent, M.; Veeraraghavan, N.; Albert, I.; Miller, W.; Makova, K. D.; Hardison, R. C.; Nekrutenko, A., A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res* **2007,** 17, (6), 960-4.

233. Kuehn, H.; Liberzon, A.; Reich, M.; Mesirov, J. P., Using GenePattern for gene expression analysis. *Curr Protoc Bioinformatics* **2008,** 22, 7.12.1-7.12.39.

234. Reich, M.; Liefeld, T.; Gould, J.; Lerner, J.; Tamayo, P.; Mesirov, J. P., GenePattern 2.0. *Nat Genet* **2006,** 38, (5), 500-1.

235. Brusniak, M. Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L. N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J. D., Corra:

Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics* **2008,** 9, 542.

Chapter 2

An Optimized Time Alignment Algorithm for LC-MS Data: Correlation Optimized Warping using Component Detection Algorithm-Selected Mass Chromatograms

Christin C, Smilde AK, Hoefsloot HC, Suits F, Bischoff R, Horvatovich PL. Anal Chem. 2008 Sep 15;80(18):7012-21

ABSTRACT

Correlation Optimized Warping (COW) based on the Total Ion Current (TIC) is a widely used time alignment algorithm (COW-TIC). This approach works successfully on chromatograms containing few compounds and having a well-defined TIC. In this paper, we have combined COW with a Component Detection Algorithm (CODA) to align LC-MS chromatograms containing thousands of biological compounds with overlapping chromatographic peaks, a situation where COW-TIC often fails. CODA is a variable selection procedure that selects mass chromatograms with low noise and low background (so-called "high-quality" mass chromatograms). High quality mass chromatograms selected in each COW segment ensure that the same compounds (based on their mass and their retention time) are used in the 2-dimensional benefit function of COW to obtain correct and optimal alignments (COW-CODA). The performance of the COW-CODA algorithm was evaluated on three types of complex datasets obtained from the LC-MS analysis of samples commonly used for biomarker discovery and compared to COW-TIC using a new global comparison method based on overlapping peak area: trypsin-digested serum obtained from cervical cancer patients, trypsin-digested serum from a single patient that was treated with varying pre-analytical parameters (factorial design study) and urine from pregnant and non-pregnant women. While COW-CODA did result in minor misalignments in rare cases, it was clearly superior to the COW-TIC algorithm, especially when applied to highly variable chromatograms (factorial design, urine). The presented algorithm thus enables automatic time alignment and accurate peak matching of multiple LC-MS datasets obtained from complex body fluids that are often used for biomarker discovery.

1 INTRODUCTION

Comparative proteomics and biomarker discovery studies often use label-free liquid chromatography coupled to mass spectrometry (LC-MS) to detect differences between pre-classified sample sets. Easily accessible body fluids such as blood (plasma or serum) and urine are used primarily for this purpose. However, analyzing body fluids is challenging since they contain a large number of diverse compounds covering a wide dynamic concentration range leading to enormous amounts of raw data that need to be processed prior to statistical comparison. LC-MS data acquired in profile mode characterize compounds by their retention time and mass to charge ratio (m/z) and the quantity is reflected in the measured intensity (e.g. ion count).

Data processing workflows must be designed in a way to extract accurate information related to the identity and quantity of the detected compounds to allow subsequent statistical analyses and to find concentration differences between pre-classified sample sets[1]. One of the most important challenges in detecting concentration differences is to ensure that identical peaks are compared across multiple samples (peak matching procedure), since liquid chromatography separations are prone to non-linear elution time shifts as a result of slight variations in flow rate, gradient slope and temperature as well as to column aging and the need to renew eluents from time to time. This is especially important for complex mixtures, such as depleted and trypsin-digested serum (shotgun proteomics approach)[2] or acid-precipitated urine, where many compounds elute with similar retention times. Improper correction of retention time shifts may thus lead to incorrect peak matching across multiple samples, resulting in statistical errors and the false discovery of biomarker candidates. Since clinical biomarker discovery and other proteomics or metabolomics applications require the comparative analysis of many samples to enhance statistical power, reliable, automatic, non-linear time alignment algorithms are required to avoid such pitfalls.

Several techniques to correct non-linear retention time shifts have been developed. These methods differ in both the search space and the benefit function (a measure of similarity) used to find the optimum retention time shift correction. Furthermore, most of the reported algorithms arbitrarily choose one chromatogram as reference and align all other "sample" chromatograms to it as a way to co-align all chromatograms.

The retention time vector in LC-MS data contains thousands of points. The search space, in which to find the optimal mapping between reference and sample retention time vectors, must be limited in order to keep computation time within reasonable limits and to avoid misalignment between distant, unrelated parts of the reference and sample chromatograms. Dynamic Time Warping (DTW)[3-6] calculates the shift of data points in two chromatographic profiles and warps the trajectories in such a way that the distance between them is minimized using a set of constraints with respect to changes that are allowed for each point of the sample retention time vector. Correlation Optimized Warping (COW)[3, 7-11] divides the chromatographic profile into segments and stretches or shrinks these in a linear manner within a limited search space to maximize the correlation to a reference chromatogram. Other approaches, such as

parametric[12] and semi-parametric warping[9] use the full length of the chromatographic profile and perform the time correction in one step. Parametric warping optimizes polynomial coefficients by minimizing the difference of the intensity for a given data point between reference and sample chromatographic profiles.

The majority of the published time alignment methods use a 1-dimensional benefit function[3, 7, 10, 12-18] to search for the optimal alignment even when the data was acquired with a detector providing 2-dimensional information in addition to the separation time (e.g. GC-MS, LC-DAD, LC-MS). For LC-MS data the benefit function is often based on the Total Ion Chromatogram (TIC, or sum of all intensities within one scan) or Base Peak Chromatogram (BPC, or the maximal intensity in each scan). Time alignment using a 1-dimensional benefit function may work well for samples where time and 1-dimensional information used for alignment are similar across the samples, but it can be inappropriate for proteomics and metabolomics samples containing a high number of partially overlapping, closely eluting peaks with varying intensities. A few time alignment methods have been reported using a 2-dimensional benefit function[5, 19-26]. Some methods specifically note the advantages of using a 2-dimensional versus a 1-dimensional benefit function[5, 24-26]. Certain methods use single-scan mass spectra, known to be noisy due to scan-to-scan fluctuations[23, 27, 28], whereas other algorithms use 2-dimensional peaks that are local maxima of the ion intensity in the retention time and m/z space[16, 19, 20, 25, 29-33].

Comparative studies have identified some of the advantages and disadvantages of warping algorithms using different search space and 1-dimensional benefit functions on samples containing a small number of well resolved compounds [3, 13, 34, 35]. However, until now no evaluation and comparison of time alignment algorithms on complex data have been reported. Bylund[10] used covariance instead of the correlation coefficient as benefit function to calculate the similarity between two chromatograms. This work concluded that the covariance measure is more sensitive to the peak height and will favor the alignment of large peaks, avoiding interference from regions containing mostly noise and background.

LC-MS chromatograms contain noise, background, and analyte peaks of varying quality with respect to the location in the chromatogram. Electrospray Ionization (ESI), which is the most often used ionization technique for LC-MS of bio-molecules, generates chemical noise and contaminants from solvents or the atmospheric environment that may be present at different parts of a chromatogram. Therefore it is necessary to examine the local information content of LC-MS data and to locate regions containing high quality information (low noise and background and a relatively high, compound-related signal) and to use those for time alignment. The Component Detection Algorithm [36, 37] (CODA) measures the information content of mass chromatograms containing a minimal amount of high-frequency noise, spikes (peak width of only one scan) and background by comparing the magnitude of change between the original trace and the mean-subtracted trace that was smoothed using a moving average. Hence, regions of high information content can be located and subsequently used for time alignment by COW.

In this chapter, we combine mass chromatogram selection using CODA with a modified COW algorithm in order to take the local information in LC-MS

chromatograms into account. The COW algorithm is applied segment-wise to pairs of selected mass chromatograms with the product correlation coefficient of the selected mass traces as 2-dimensional benefit function. The performance of the COW-CODA algorithm was evaluated using LC-MS data of urine and trypsin-digested human serum obtained from real-case proteomics and metabolomics studies[38-40]. These datasets exhibit different degrees of non-linear retention time shifts and contain a large number of compounds of highly variable properties and amounts.

2 THEORY

2.1 Conditions for proper time alignment using COW-CODA

A number of conditions have to be met for successful application of the COW-CODA algorithm. Alignment is based on the presence of common peaks (compounds) between the reference and the sample chromatograms. The first criterion is thus to find a minimal number of common peaks in each time segment that should be aligned using the COW algorithm. In case there are no high-quality mass traces in a given segment, this segment is left unchanged, as there is no information to base the alignment on. This should, however, be the exception, if an overall well-aligned dataset is to be obtained. The criterion of finding common, high-quality mass chromatograms is generally satisfied for all easily accessible body fluids in areas where compounds (peptides, metabolites) elute, since even highly variable body fluids such as urine contain a large number of conserved compounds.

We selected COW as the search algorithm for time alignment, because it corrects non-linear time shifts by stretching or shrinking the data in a segment-wise fashion until optimal correlation has been reached. The limited search space of retention time range reduces the risk of large retention time corrections resulting in erroneous time alignments often occurring when the search space is too large. The aim of combining COW with CODA was fourfold: 1. assure that peaks with similar retention times but different m/z values are considered as separate features in the benefit function, 2. consider only data from common peaks between the reference and sample datasets in the benefit function, 3. avoid traces containing high noise and background, and 4. take into account that peaks, background and noise distribution vary strongly between different regions of the LC-MS dataset.

2.2 Component Detection by CODA

The CODA algorithm[36, 37] was developed to select mass chromatograms with high quality peaks, low noise and low background. This algorithm contains two main steps: detection of spikes (single scan signals originating from electronic noise) and detection of high signal background (typically originating from the mobile phase). When a mass chromatogram contains noise and spikes, the smoothed version will be different from the original chromatogram. We use a moving average to smooth the data in a specific m/z trace segment using a window larger than the peak width of the spikes, which results in large differences between the smoothed and original data when spikes are present and a low CODA similarity index. A mass chromatogram that has a high

level of background noise will have a relatively high mean value. Hence it will differ strongly from its mean-subtracted version resulting also in a low similarity index. In contrast, a mass chromatogram with no spikes, a low level of background noise and significant peaks will have a high similarity to its mean-subtracted version. The quality of a mass chromatogram is defined by a single, combined index of the two similarity indices. A high-similarity index thus indicates high quality mass chromatograms with intense peaks. This similarity index is called Mass Chromatographic Quality (*MCQ*) with a minimum value of 0 and a maximum value of 1. The CODA algorithm has been described in detail by Windig *et al* .

2.3 Combining COW and CODA (COW-CODA)

2.3.1 Segmentation and search space

The COW algorithm, as described by Nielsen *et al*.[7], was employed in our procedure, with a modification of the benefit function, which was originally based on the sum of correlations of each segment using 1-dimensional information (e.g. TIC or single wavelength UV traces). Assume that we want to align a sample chromatogram **S** to a reference chromatogram **R**, with the number of scans in the reference and sample chromatograms being L^S and L^R, respectively. Each chromatogram may vary in length due to the different number of mass spectrometric scans (especially for ion trapping instruments). In our case, the number of mass traces d in each chromatogram is equal since we have used the same mass range for data acquisition (100-1500 Da) and both chromatograms have been identically smoothed from their original 0.1 amu resolution to 1 amu (see Gaussian smoothing and data reduction in Material and Methods). In the COW algorithm, the reference chromatogram remains unchanged while the endpoints of the sample chromatogram segments are allowed to move according to three constraints (1) start and (2) endpoints of the sample chromatograms are unchanged, and (3) sample segments lengths are allowed to change with the slack parameter (see point 3 of the Theory part in the support information). The flowchart of the warping process for two chromatograms is schematically described in Figure 1, while the detailed steps of the warping algorithm are described in the supporting information. COW specifies the degree to which the segment endpoints may move through the "slack" parameters. Details about the search space given by the slack parameters are discussed in Nielsen et al[7]. and in the Theory part of the supporting information.

2.3.2 Segment-wise mass chromatogram selection

The procedure for selecting high quality mass chromatogram is the same for each segment. The *MCQ* values from the CODA algorithm were calculated for each mass chromatogram from a given segment of chromatogram **S** and **R**, resulting in two vectors of length d containing the *MCQ* values of the corresponding traces. Since the segment endpoints of the reference chromatograms are fixed and the segments endpoints of the sample chromatograms can vary, a larger segment is chosen for the sample chromatogram compared to the corresponding segment in the reference chromatogram (a detailed description how segment endpoints of the mass chromatograms are obtained

is given under point 6 of the Theory part in the supporting information). The product of the *MCQ* values of the corresponding mass chromatograms for each segment of the sample and reference chromatograms were calculated and used as a measure of quality. Parameters were set to select mass chromatograms with *MCQ* products higher than 0.59 with an upper limit of 30 mass chromatograms per segment.

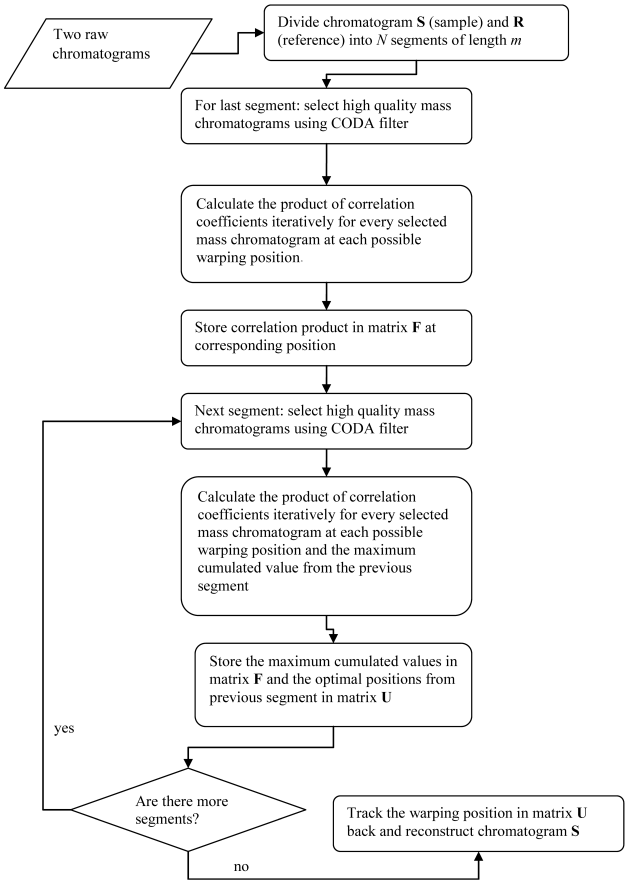

Figure 1. Flowchart of the COW-CODA algorithm, showing the main steps in warping a sample chromatogram **S** to the reference chromatogram **R**. This process is repeated until all sample chromatograms have been aligned to the chosen reference chromatogram in the dataset.

For a given segment the product of correlation coefficients was calculated for each pair of selected mass chromatograms using the allowed segment endpoints for the reference and the sample chromatograms (see point 3 of the Theory section in the supporting information). The warping procedure uses dynamic programming. In case there is no mass trace selected for a segment, because there is no high quality mass chromatogram present, no warping is applied for this segment and the cumulated correlation from the previous segment is passed on to the next segment without modification. The algorithm may be improved by requesting a minimum number of selected mass traces for warping and relating the number of selected traces and the

product of their *MCQ* values to the allowed retention time correction thus tolerating larger retention time shifts for segments containing a high number of high-quality mass traces.

The segment wise trace selection using *MCQ* products enables not only the selection of high-quality mass traces containing information about the peaks but also the selection of traces from peaks occurring in both chromatograms. This favors alignments that are based on conserved peaks (compounds) and reduces the risk of misalignments in crowded regions of chromatograms from highly complex samples.

2.3.3 Form of the benefit function

The benefit function in COW using 1-dimensional information for time alignment is the sum of the Pearson correlation between segments of reference and sample chromatograms. In COW-CODA, this output function is replaced with the correlation product of segments for the selected mass traces. The overall benefit function is therefore the sum of the correlation products. A detailed description of how the correlation product and the corresponding benefit function are obtained is given in Figure S-1 (supporting information).

2.4 Choosing the reference chromatogram

Aligning multiple chromatograms requires selection of a reference chromatogram to which the other chromatograms in the dataset shall be aligned. To select the best reference, we have calculated the correlation of LC-MS chromatograms based on the reconstructed TIC from all CODA-selected mass traces (CODA-TIC) with 100 different *MCQ* thresholds between 0.80-0.99 (with an interval of 0.0019) across the entire retention time range, during which relevant peaks elute (~44-130 minutes for serum datasets and ~13-105 minutes for the urine dataset). For each chromatogram and each studied *MCQ* value, we calculated the sum of the correlation coefficients of the CODA-TIC profile before time alignment and the chromatogram giving the highest sum of correlation was chosen as the best reference. Consequently, the chromatogram giving the lowest sum was chosen as the worst reference. The selected best and worst reference chromatograms were plotted as a function of *MCQ* value and the chromatograms that were most often selected as best and worst references were chosen to perform the time alignment giving a "best" and "worst case" scenario. The complete flowchart of reference chromatogram selection is presented in Figure S-2 (supporting information).

2.5 Global evaluation of the time alignment quality

We can easily compare the quality of time alignment for single peaks by visual inspection of multiple datasets using internal standards (peptides or metabolites) that were added to the samples. However, it is difficult to judge the overall quality of the time warping algorithm for complex chromatograms by visualizing only a small number of selected compounds. To overcome this limitation, we have developed a procedure to assess the quality of time alignment based on the calculation of the overlapping peak area between pairs of chromatograms. Peak areas were calculated after applying a

modified local baseline subtraction method called M-N rules using M=3 and N=8[22]. An increased quality of time alignment between two chromatograms is reflected in a larger overlapping peak area. The sum of overlapping peak areas from each chromatogram to all of the other chromatograms in the original dataset is then compared to the overlapping peak area after applying COW-TIC or COW-CODA. Using this approach, the performance of the time alignment methods can be compared relative to each other. This evaluation method considers a much larger number of chromatographic peaks than the internal standard method. It is, however, not able to judge if the time alignment method is entirely accurate. In order to check the number of possible misalignments after warping, we visually inspected three peaks per time segment (number of segments were 51, 84 and 41 for Dataset 1, 2 and 3 respectively) having the largest average peak intensity in each dataset.

3 MATERIAL AND METHODS

3.1 Chemicals

Acetonitrile (ACN) HPLC-S gradient grade (Biosolve; Valkenswaard, The Netherlands), ultra pure water (18.2 MΩ/cm), trifluoroacetic acid (TFA) 99% spectrophotometric grade (Aldrich; Milwaukee, USA) and formic acid (FA) 98-100% pro analysis (Merck; Darmstadt, Germany) were used for reagent and solvent preparation.

3.2 Serum samples

3.2.1 Cervical cancer patients (Dataset 1)

Serum samples from 10 cervical cancer patients at two different time points (time point A: before treatment with confirmed cancer; time point B: after treatment with no recurrence of the cancer for at least 6 months) were obtained from the Department of Gynecological Oncology (University Medical Centre Groningen, The Netherlands). The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004. All samples were stored at –80°C in aliquots. The samples from cervical cancer patients were depleted of the 6 most abundant serum proteins on a Multiple Affinity Removal column (4.6 × 50 mm, # 5185-5984, Agilent Technologies) followed by digestion of the remaining proteins with trypsin. Further details about the LC-MS analyses are described in Govorukhina *et al*[40].

3.2.2 Factorial design (Dataset 2)

For the factorial design study serum samples from a healthy female volunteer were obtained from the Department of Gynecological Oncology (University Medical Centre Groningen, The Netherlands) and stored at –80°C in aliquots. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

The sample preparation of the serum in this dataset was similar to the sample preparation of the serum for dataset 1, except for the seven following pre-analytical

parameters, which were varied at two levels according to a 2_{IV}^{7-3} fractional factorial design (see Table 1). In brief, these factors were type of blood collection tube, hemolysis level, clotting time, number of freeze-thaw cycles, trypsin concentration, deactivation of trypsin after digestion and stability of the digested sample in the autosampler of the LC-MS system. 16 from the 128 possible combinations were selected with four repetitions (chromatogram number 1,4,11,14) of one condition resulting in a total number of 19 analyses. Digested serum samples were analyzed by LC-MS according to Govorukhina *et al.*[40].

Exp. Name	Run order	Blood collection tube	Hemo-Lysis	Clotting Time (hour)	Freeze-thaw Cycles	Trypsin digestion	Stopping trypsin	Stability sample (days)
N1	1	BD368430	Low	2	1 cycle	1:20	Yes	0
N2	3	BD367784	Low	2	1 cycle	1:100	Yes	0
N3	9	BD368430	High	2	1 cycle	1:100	No	0
N4	17	BD367784	High	2	1 cycle	1:20	No	30
N5	18	BD368430	Low	6	1 cycle	1:100	No	30
N6	2	BD367784	Low	6	1 cycle	1:20	No	0
N7	10	BD368430	High	6	1 cycle	1:20	Yes	30
N8	16	BD367784	High	6	1 cycle	1:100	Yes	0
N9	8	BD368430	Low	2	3 cycles	1:20	No	30
N10	15	BD367784	Low	2	3 cycles	1:100	No	0
N11	13	BD368430	High	2	3 cycles	1:100	Yes	30
N12	7	BD367784	High	2	3 cycles	1:20	Yes	0
N13	12	BD368430	Low	6	3 cycles	1:100	Yes	0
N14	6	BD367784	Low	6	3 cycles	1:20	Yes	30
N15	5	BD368430	High	6	3 cycles	1:20	No	0
N16	9	BD367784	High	6	3 cycles	1:100	No	30
N17	14	BD368430	Low	2	1 cycle	1:20	Yes	0
N18	1	BD368430	Low	2	1 cycle	1:20	Yes	0
N19	4	BD368430	Low	2	1 cycle	1:20	Yes	0

Table 1. Overview over pre-analytical parameters and their levels in a fractional factorial design study of depleted and trypsin-digest serum (Dataset 2).

3.3 Urine samples (Dataset 3)

First-void midstream morning urine samples were obtained from 25 pregnant females from a local biobank (Department of Obstetrics and Gynecology of the University Medical Center in Groningen, The Netherlands) and stored frozen at -20 ºC. Twenty-five first-void midstream morning urine samples from non-pregnant females were collected in polypropylene containers and kept at 4 °C for a maximum of 1 day, after which they were aliquoted in 10 ml polypropylene tubes and stored at -20 °C. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

Urine samples were thawed, mixed and acidified with TFA to reach a final concentration of 1%. Samples were left overnight on melting ice, and centrifuged to

remove precipitate (10 min at 1500 g and 4 °C). The supernatant was diluted 1:1 with 0.2% FA in 10% ACN and stored at 4 °C until analysis. Supernatants of the acid-precipitated urine samples were analyzed by LC-MS as described by Kemperman *et al.*[39].

3.4 *Data Analysis*

For processing and multivariate statistical analysis the original Bruker Daltoniks LC-MS data files were converted to ASCII-format with the Bruker Data Analysis software. The ASCII files were transformed into a matrix with the dimensions: retention time, m/z value and intensity. Data reduction was performed to combine m/z ratios into 1 amu bins (originally 0.1 amu) by multiplying the original data with a weight-normalized two-dimensional Gaussian weight matrix. This was followed by time alignment using the COW-CODA or the original COW-TIC algorithm. All alignments were done with respect to the best and the worst reference chromatogram. The following parameters were used for the datasets obtained from analyzing trypsin-digested serum (factorial design and cervical cancer respectively) segment length *m*: 139 data points (~2.3 min) and 84 data points (1.5 min), dividing each chromatogram into 51 and 84 segments; slack parameter *t*: 28 and 17 data points. The following parameters were used for the dataset obtained from analyzing acid-precipitated urine: segment length *m*: 83 data points (~2.2 min), number of segments: 41; slack parameter *t* 16 data points.

For each aligned chromatogram, a peak list was generated using M-N rules with M set to 3 and N to 8. The peak lists, generated from all chromatograms (samples), were used to create one common matched peak matrix per study. In order to combine peak lists, one-dimensional peak matching was performed using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks 0.75 min). Missing peak locations were filled with the calculated, background-subtracted local intensity in the respective chromatogram obtained at a location determined from the average m/z and retention time of the corresponding peaks in samples where they were detected. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix.

All data preprocessing work was done on a personal computer equipped with a +3800 MHz AMD processor and 4 GB of RAM. The Matlab code of this software is available at https://trac.nbic.nl/lcms_time_alignment_algorithms/.

4 RESULTS AND DISCUSSION

4.1 *Design of the study*

The time alignment algorithm was applied and evaluated with three different datasets obtained from proteomics or metabolomics profiling studies. The first set of analyzed samples was serum depleted of the 6 most abundant proteins and trypsin-digested. These samples resulted from a study to discover novel biomarker candidates for cervical cancer and they are typical for highly complex, shotgun proteomics samples. Previous work had shown that the concentration sensitivity of this method lies in the

0.5µM range[40], an area where mainly high-abundant proteins with low biological (inter-patient) variability are detected. It is, however, noteworthy that a number of investigators have recently shown that also this part of the proteome may change in a disease-dependent manner due to residual proteolytic activity[41]. This LC-MS dataset (referred to as Dataset 1) was acquired under strict standard operating conditions and thus contained low analytical variability[40]. The factorial design dataset (Dataset 2) was obtained from depleted and trypsin-digested serum from one healthy female volunteer and thus contained no biological variability. However, seven pre-analytical factors were deliberately varied and the effect on the overall proteomics profile studied at two different levels resulting in considerable analytical variability (see Table 1 for details). Dataset 3 was obtained by analyzing acid-precipitated urine, a body fluid containing largely low-molecular weight metabolic end products of the organism destined for excretion, from 50 different women. Due to a very high level of biological variability in this dataset, the generation of a common aligned peak matrix from different LC-MS analyses proved particularly challenging. As this dataset was also acquired under stringent standard conditions, it had low analytical variability[39]. The serum samples contain on average 10800 features extracted under the specified MN rules, corresponding to 2200-3600 peaks, while the urine samples contain 13550 extracted features, corresponding to 2700-4500 peaks. The distribution of peaks in the retention time-m/z space is quite uniform for urine samples, while the serum samples show an elliptical distribution from low retention time and low m/z values to high retention time and high m/z values (see Figure S-3 in the supporting information).

4.2 Comparison between the COW-TIC and the COW-CODA algorithm

The performance of the COW-TIC and the COW-CODA algorithms was evaluated based on annotated peaks originating from added, known compounds as well as by using a global evaluation strategy based on the calculated overlapping peak areas between the reference and the various sample chromatograms.

4.2.1 Evaluation based on added, known compounds

Figure 2 shows extracted ion chromatograms of selected internal standards using the original data (black traces), aligned data after applying the COW-TIC (blue traces) or the COW-CODA algorithm (red traces) using the best reference chromatograms. Visualization of these peaks (spiked peptides for urine samples and horse heart Cytochrome C-derived tryptic peptide fragments for serum samples) in the original datasets shows that the initial retention time shifts are on the average 0.52 min for Dataset 1 (cervical cancer), 0.23 min for Dataset 2 (factorial design) and 0.31 min for Dataset 3 (urine). The COW-TIC algorithm using a 1-dimensional benefit function was only able to align Dataset 1, which produced well-defined, highly similar TICs across all samples despite the sometimes rather large differences in retention time (up to 2.5 min) between runs (Figure 2a). Results obtained with the COW-CODA algorithm were comparable to COW-TIC in this case.

Although the complexity of the original Dataset 2 is similar to Dataset 1, the COW-TIC algorithm was unable to align the chromatograms correctly due to rather dissimilar TIC profiles (large analytical variability due to the deliberate variation of pre-analytical parameters). Indeed, in some chromatograms COW-TIC increased the retention time shift differences compared to the original time differences. The COW-CODA algorithm, on the other hand, resulted in clearly improved alignment resembling the result obtained with Dataset 1 (Figure 2b). The difference in performance of the algorithms was even more pronounced when trying to align the 50 chromatograms of Dataset 3 (urine), which contains large biological variability. The COW-TIC algorithm (Figure 2c) lead to a number of obvious misalignments, while COW-CODA resolved the initial misalignments correctly leading to a well-matched set of data, except for the standard peak eluting around 39 min (see Figure 2c). Even for this peak retention time shifts were considerably reduced. Thus only application of the COW-CODA algorithm resulted in clearly improved alignments of all studied datasets. Similar results were obtained when aligning to the worst reference chromatogram (Figure S-4 in the supporting information) indicating that selection of the reference chromatogram has no effect on the final quality of warping. This simplifies data processing, since one may start the alignment with any chromatogram as the reference.

Performance of the COW-CODA algorithm might be improved by applying a higher extent of smoothing on a separate copy of the LC-MS data to be used for calculating the benefit function, while segment-based mass trace selection using CODA is still performed on the original LC-MS data using a small extent of smoothing or no smoothing at all. However, time alignment with COW-CODA was sufficiently accurate for correct peak matching.

Representing the overlapping peak area of the reference and sample chromatograms per segment reveals the performance differences of the two time alignment algorithms for different segments. Figure S-6 shows the overlapping peak area of the segments between the best reference and a given sample chromatogram. The figure shows that for most of the segments the two algorithms give similar results. While COW-TIC is performing slightly but not substantially better than COW-CODA for some segments (segments 16-18, 20, 22-23 and 28 show slightly better performance and segment 19 shows moderately better performance), COW-CODA improves alignment of segments 21, 25-27, 29-31, 33, 38 with substantial improvements for segments 21, 29 and 33.

Retention time correction as a function of retention time obtained with the COW-CODA and COW-TIC algorithm are presented in Figure S-7 (Supporting Information; for chromatograms 1 and 14 for Dataset 2 (Factorial design)) showing substantial corrections between retention times of 110 and 125 min and the superior performance of the COW-CODA algorithm in this difficult region (see Figure S-7).

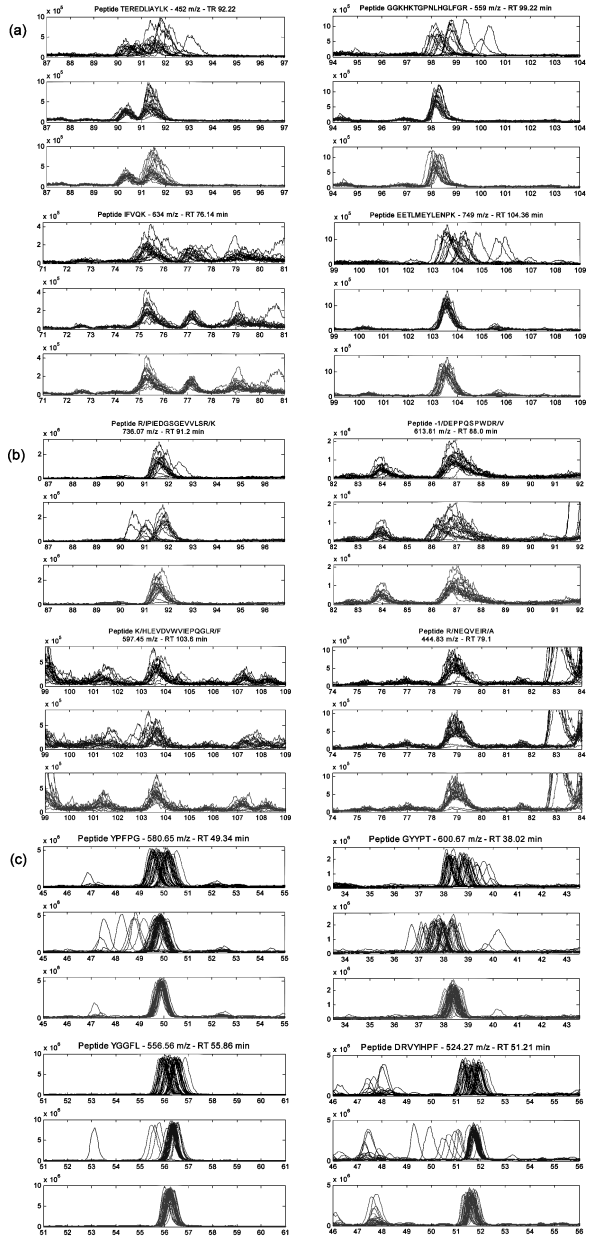

Figure 2. Extracted ion chromatograms of internal standard peptides (a) Dataset 1 (cervical cancer serum) comprised of 20 chromatograms, (b) Dataset 2 (factorial design serum) comprised of 19 chromatograms, and (c) Dataset 3 (acid-precipitated urine) comprised of 50 chromatograms. Each peptide is presented before alignment (top/black), after alignment by COW-TIC (middle/blue), and after alignment by COW-CODA (bottom/red). These time alignment results were obtained using the best references, which were chromatograms number 2, 14 and 25 for Datasets 1, 2 and 3, respectively. Significant misalignments are observed for Datasets 2 and 3 when using the COW-TIC algorithm, while the COW-CODA algorithm resulted in well-aligned peak clusters.

4.2.2 Global evaluation of alignment quality

The sums of overlapping peak areas between pairs of chromatograms from the same dataset were obtained with the original data, after applying the COW-TIC or the COW-CODA algorithm, using the best or the worst reference chromatogram (Figure 3 and Figure S-5 in SI). The COW-CODA algorithm improved time alignment for all three datasets compared to the original, non-aligned data and for Datasets 2 and 3 with respect to the COW-TIC algorithm confirming results obtained with the internal standards. The COW-TIC algorithm led to higher overlapping peak areas than COW-CODA for Dataset 1, but the difference between the performances of the two algorithms is very small. The slightly better performance of COW-TIC in this case could be due to the fact that CODA-selected single mass traces contain higher levels of noise than the TIC, where the noise of the individual mass traces averages out. The slightly better time alignment using COW-TIC was also observed when inspecting individual peaks of tryptic peptides derived from the added Cytochrome C (see Figure 2a).

4.3 Effect of reference chromatograms

It may be argued that all COW-based algorithms depend strongly on the selected reference chromatogram (see Figure S-8 of the supporting information and the Material and Methods section on how reference chromatograms were selected). In order to study this factor in more detail, we compared the overlapping peak areas (Figure S-5) confirming earlier results that showed that the overall quality of alignment using the CODA-COW algorithm depends little on the reference chromatogram. However, small differences were observed in Dataset 1 for chromatograms 18 and 19 and for a few chromatograms in Dataset 3 (Figures S-5a and c). The stability of the algorithm for Datasets 1 and 2 with respect to the choice of the reference chromatogram increases our confidence that the extent of misalignment with any of the chosen reference chromatograms is small. We assume that a large number of misalignments would have resulted in variable overlapping peak areas depending on the selected reference chromatogram. The fact that the COW-CODA algorithm functions correctly and independently of the chosen reference chromatogram greatly simplifies automation of the time alignment step as part of the overall data processing pipeline for chromatograms obtained from depleted, trypsin-digested serum samples. The slight dependency of the final alignment on the chosen reference chromatogram for Dataset 3 shows, however, the need for algorithms that select the best reference for datasets with high biological variability. Calculating the overlapping peak area appears to be a suitable way to do so.

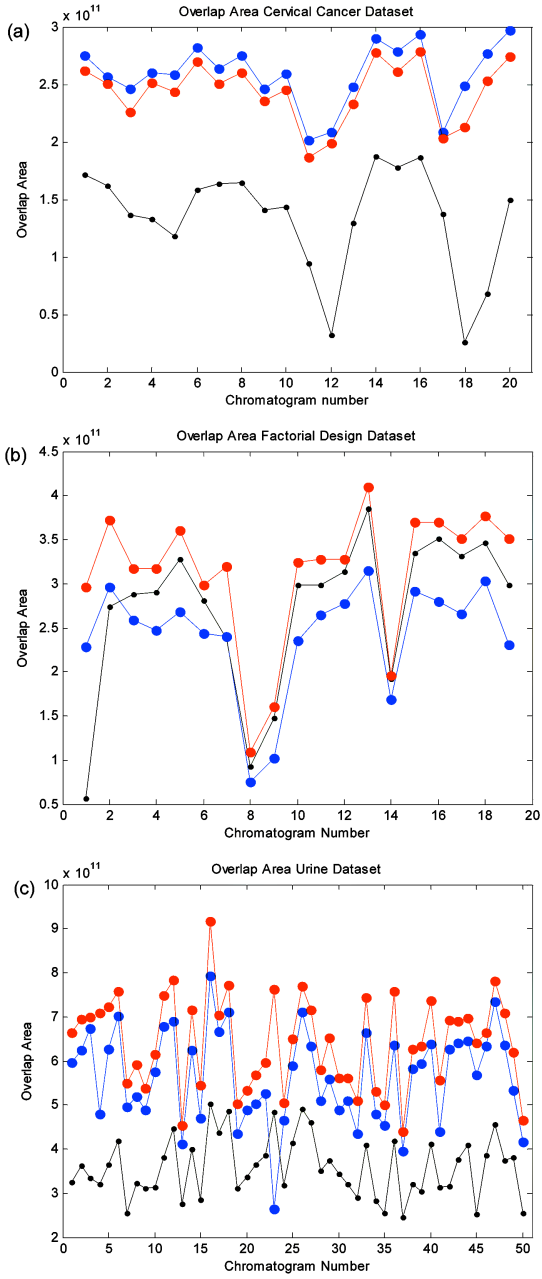

Figure 3. Calculated overlapping peak area after application of M-N rules to (a) Dataset 1 (cervical cancer serum), (b) Dataset 2 (factorial design serum), and (c) Dataset 3 (acid-precipitated urine). Original non-aligned data (black), aligned with the COW-TIC (blue) or with the COW-CODA algorithm (red). COW-CODA (red) improved the alignment for all three datasets compared to the non-aligned data and with respect to COW-TIC for Datasets 2 and 3. These time alignment results were obtained using the best references, which was the chromatogram number 2, 14 and 25 for Datasets 1, 2 and 3, respectively.

4.4 *Assessing the inherent variability of the datasets and the required processing time*

With a properly aligned common peak matrix, it is possible to assess the variability of the different datasets. A first evaluation of the variability of Datasets 1-3 was performed by plotting the number of detected peaks against the relative standard deviation (RSD)(Figure 4). Based on considerations discussed before, it is expected that the relative standard deviation (RSD) histogram obtained from the aligned peak matrix of the depleted and trypsin-digested serum samples from cervical cancer patients (Dataset 1) would be narrower relative to histograms obtained from the factorial design or urine datasets. Figure 4 shows further that the biological variability of urine not only increases the maximum RSD, but results also in a broader RSD distribution compared to serum datasets, indicating that peaks have rather high variability in intensity in urine. This stands in contrast to the serum-derived Dataset 2, where variation is only due to pre-analytical parameters, which increases only the mode of RSD without broadening the RSD distribution. Accurate time alignment thus allows to measure the variability that is inherent to the datasets resulting from analytical and biological variations. Such histograms may be used to perform statistical simulations, e.g. for power calculations, or to test the performance of different variable selection and classification algorithms.

Alignment of one pair of chromatogram obtained with serum samples (~7100 scans) took ~22 minutes using COW-CODA and 0.8 minutes using COW-TIC with an ordinary PC as described in the Material & Methods section. The processing time for chromatograms obtained from urine samples (~3450 scans) was 5.35 minutes for COW-CODA and 0.2 minutes for COW-TIC. The ~27-times higher processing time of COW-CODA is mainly due to the higher I/O, since the algorithm operates on the full LC-MS dataset contrary to the COW-TIC algorithm, which is only using a single trace, the TIC, for the time alignment procedure.

5 CONCLUSIONS

We describe an improved time alignment algorithm based on COW combined with a segment-wise selection of high-quality mass chromatograms using CODA. This algorithm is effective in aligning highly complex proteomics and metabolomics LC-MS datasets containing different levels of analytical and biological variability. The novel algorithm outperforms the original COW-TIC algorithm in the case of datasets containing either a high level of analytical (factorial design study) or biological (acid-precipitated urine) variability.

The presented algorithm uses a 2-dimensional benefit function in order to discriminate between peaks with different m/z values eluting at similar or identical retention times and to select mass traces sharing common, high-quality peaks. We have observed only very minor misalignments after visually inspecting a few hundred Extracted Ion Chromatograms for each of the analyzed samples. This strongly supports the conclusion that the COW-CODA algorithm results in a very high quality of time alignment and that the studied datasets contain a sufficient number of common peaks in each time segment to drive the alignment procedure to a global optimum.

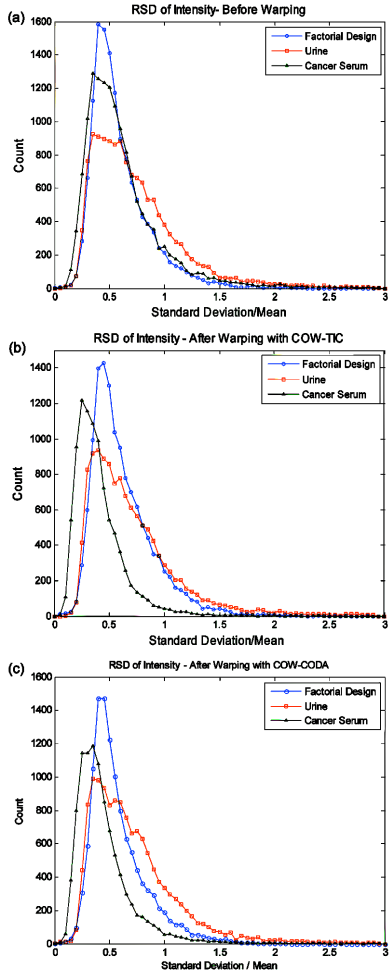

Figure 4. Standard deviation of the intensity of peaks selected with M-N rules (M = 3; N = 8) in Datasets 1 (cervical cancer serum) (green), 2 (factorial design serum) (blue) and 3 (acid-precipitated urine) (red) divided by the mean before warping (top) and after warping with the COW-CODA algorithm (middle) and after warping with COW-TIC (bottom). These time alignment results were obtained using the best references, which were chromatograms 2, 14 and 25 for Datasets 1, 2 and 3, respectively.

Another advantage of COW-CODA is that only a few parameters need to be selected and optimized depending on the dataset (e.g. adapting the chosen segment length to the observed chromatographic peak width) and that the final quality of alignment is rather independent on the initially chosen reference chromatogram.

This chapter present results from low-resolution MS data. However, COW-CODA is also applicable to high-resolution data using Gaussian smoothing and data reduction to 1 amu in the mass dimension, since the algorithm only calculates the corrected retention time of the mass scans (Figure S-9 in SI). This "binning" does not

affect the ultimate resolution of the MS data, since the warped retention time values of the mass scans can be applied to data with the original resolution.

6 REFERENCES

1. Horvatovich, P.; Govorukhina, N.; Bischoff, R., Biomarker discovery by proteomics: challenges not only for the analytical chemist. *Analyst* **2006,** 131, (11), 1193-6.
2. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd, Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **1999,** 17, (7), 676-82.
3. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* **2004,** 18, (5), 231-241.
4. Ramaker, H.-J.; van Sprang, E. N. M.; Westerhuis, J. A.; Smilde, A. K., Dynamic time warping of spectroscopic BATCH data. *Anal Chim Acta* **2003,** 498, 133-153.
5. Prince, J. T.; Marcotte, E. M., Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal Chem* **2006,** 78, (17), 6140-6152.
6. Jaitly, N.; Monroe, M. E.; Petyuk, V. A.; Clauss, T. R.; Adkins, J. N.; Smith, R. D., Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal Chem* **2006,** 78, (21), 7397-409.
7. Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. r., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J Chromatogr A* **1998,** 805, 17-35.
8. Fransson, M.; Folestad, S., Real-time alignment of batch process data using COW for on-line process monitoring. *Chemometrics and Intelligent Laboratory Systems* **2006,** 84, (1-2), 56-61.
9. van Nederkassel, A. M.; Xu, C. J.; Lancelin, P.; Sarraf, M.; MacKenzie, D. A.; Walton, N. J.; Bensaid, F.; Lees, M.; Martin, G. J.; Desmurs, J. R.; Massart, D. L.; Smeyers-Verbeke, J.; Vander Heyden, Y., Chemometric treatment of vanillin fingerprint chromatograms: Effect of different signal alignments on principal component analysis plots. *Journal of Chromatography A* **2006,** 1120, (1-2), 291-298.
10. Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *Journal of Chromatography A* **2002,** 961, (2), 237-244.
11. Sadygov, R. G.; Martin Maroto, F.; Huhmer, A. F. R., ChromAlign: A Two-Step Algorithmic Procedure for Time Alignment of Three-Dimensional LC,àíMS Chromatographic Surfaces. *Analytical Chemistry* **2006,** 78, (24), 8207-8217.
12. Eilers, P. H. C., Parametric time warping. *Anal Chem* **2004,** 76, (2), 404-411.
13. Pravdova, V.; Walczak, B.; Massart, D. L., A comparison of two algorithms for warping of analytical signals. *Anal Chim Acta* **2002,** 456, 77-92.
14. Malmquist, G.; Danielsson, R., Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *Journal of Chromatography A* **1994,** 687, (1), 71-88.
15. Kassidas, A.; MacGregor, J. F.; Taylor, P. A., Synchronization of batch trajectories using dynamic time warping. *AIChE Journal* **1998,** 44, (4), 864-875.
16. Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E., High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *Journal of Chromatography A* **2003,** 996, (1-2), 141-155.
17. Listgarten J.; Neal R.M.; Roweis S.T.; A., E., Multiple alignment of continuous time series. *Advances in Neural Information Processing Systems Cambridge, MA* **2005,** 17.
18. Nordström, A.; O'Maille, G.; Qin, C.; Siuzdak, G., Nonlinear Data Alignment for UPLC-MS and HPLC-MS Based Metabolomics:‚Äâ Quantitative Analysis of Endogenous and Exogenous Metabolites in Human Serum. *Analytical Chemistry* **2006,** 78, (10), 3289-3295.
19. Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M., A suite of

algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **2006,** 22, (15), 1902-1909.

20. Li, X.-j.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R., A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry. *Molecular & Cellular Proteomics* **2005,** 4, (9), 1328-1340.
21. Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B., Signal Maps for Mass Spectrometry-based Comparative Proteomics. *Molecular & Cellular Proteomics* **2006,** 5, (3), 423-432.
22. Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A., Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* **2004,** 3, (10), 984-997.
23. Wang, P.; Tang, H.; Fitzgibbon, M. P.; McIntosh, M.; Coram, M.; Zhang, H.; Yi, E.; Aebersold, R., A statistical method for chromatographic alignment of LC-MS data. *Biostatistics* **2007,** 8, (2), 357-67.
24. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Analytical Chemistry* **2003,** 75, (18), 4818-4826.
25. Fischer, B.; Grossmann, J.; Roth, V.; Gruissem, W.; Baginsky, S.; Buhmann, J. M., Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* **2006,** 22, (14), e132-40.
26. Kirchner, M., Saussen, B., Steen, H., Steen, J., Hamprecht, J., amsrpm: Robust Point Matching for Retention Time Aligment of LC/MS Data with R. *Journal of Statistical Software* **2007,** 18, (4).
27. Piening, B. D.; Wang, P.; Bangur, C. S.; Whiteaker, J.; Zhang, H.; Feng, L. C.; Keane, J. F.; Eng, J. K.; Tang, H.; Prakash, A.; McIntosh, M. W.; Paulovich, A., Quality control metrics for LC-MS feature detection tools demonstrated on Saccharomyces cerevisiae proteomic profiles. *J Proteome Res* **2006,** 5, (7), 1527-34.
28. Finney, G. L.; Blackler, A. R.; Hoopmann, M. R.; Canterbury, J. D.; Wu, C. C.; MacCoss, M. J., Label-Free Comparative Analysis of Proteomics Mixtures Using Chromatographic Alignment of High-Resolution Œº LC‚àíMS Data. *Analytical Chemistry* **2008,** 80, (4), 961-971.
29. Katajamaa, M.; Oresic, M., Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* **2005,** 6, 179.
30. Johnson, K. J.; Prazen, B. J.; Young, D. C.; Synovec, R. E., Quantification of naphthalenes in jet fuel with GC x GC/Tri-PLS and windowed rank minimization retention time alignment. *J Sep Sci* **2004,** 27, (5-6), 410-6.
31. Johnson, K. L.; Mason, C. J.; Muddiman, D. C.; Eckel, J. E., Analysis of the low molecular weight fraction of serum by LC-dual ESI-FT-ICR mass spectrometry: precision of retention time, mass, and ion abundance. *Anal Chem* **2004,** 76, (17), 5097-103.
32. H. Robert Bergen, III; George, V.; William, A. C.; Kenneth, L. J.; Ann, L. O.; David, C. M., Discovery of ovarian cancer biomarkers in serum using NanoLC electrospray ionization TOF and FT-ICR mass spectrometry. *Disease Markers* **2004,** 19, (4), 239-249.
33. Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R., Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. *J Am Soc Mass Spectrom* **2007,** 18, (10), 1835-43.
34. van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H.; Heyden, Y. V., A comparison of three algorithms for chromatograms alignment. *J Chromatogr A* **2006,** 1118, (2), 199-210.
35. Szymanska, E.; Markuszewski, M. J.; Capron, X.; van Nederkassel, A. M.; Vander Heyden, Y.; Markuszewski, M.; Krajka, K.; Kaliszan, R., Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides. *Electrophoresis* **2007,** 28, (16), 2861-73.
36. Windig, W.; Smith, W. F., Chemometric analysis of complex hyphenated data. Improvements of the component detection algorithm. *J Chromatogr A* **2007,** 1158, (1-2), 251-7.
37. Windig, W.; Phalp, J. M.; Payne, A. W., A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry* **1996,** 68, (20), 3602-3606.

38. Horvatovich, P.; Govorukhina, N. I.; Reijmers, T. H.; van der Zee, A. G.; Suits, F.; Bischoff, R., Chip-LC-MS for label-free profiling of human serum. *Electrophoresis* **2007,** 28, (23), 4493-505.

39. Kemperman, R. F.; Horvatovich, P. L.; Hoekman, B.; Reijmers, T. H.; Muskiet, F. A.; Bischoff, R., Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: method development, evaluation, and application to proteinuria. *J Proteome Res* **2007,** 6, (1), 194-206.

40. Govorukhina, N. I.; Reijmers, T. H.; Nyangoma, S. O.; van der Zee, A. G.; Jansen, R. C.; Bischoff, R., Analysis of human serum by liquid chromatography-mass spectrometry: improved sample preparation and data analysis. *J Chromatogr A* **2006,** 1120, (1-2), 142-50.

41. Villanueva, J.; Shaffer, D. R.; Philip, J.; Chaparro, C. A.; Erdjument-Bromage, H.; Olshen, A. B.; Fleisher, M.; Lilja, H.; Brogi, E.; Boyd, J.; Sanchez-Carbayo, M.; Holland, E. C.; Cordon-Cardo, C.; Scher, H. I.; Tempst, P., Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest* **2006,** 116, (1), 271-84.

Chapter 3

Time Alignment Algorithms based on Selected Mass Traces for Complex LC-MS Data

Christin C, Hoefsloot HC, Smilde AK, Suits F, Bischoff R, Horvatovich PL. Journal Proteome Res. 2010 Mar 5;9(3):1483-95.

ABSTRACT

Time alignment of complex LC-MS data remains a challenge in proteomics and metabolomics studies. This work describes modifications of the Dynamic Time Warping (DTW) and the Parametric Time Warping (PTW) algorithms that improve the alignment quality for complex, highly variable LC-MS data sets. Regular DTW or PTW use one-dimensional profiles such as the Total Ion Chromatogram (TIC) or Base Peak Chromatogram (BPC) resulting in correct alignment if the signals have a relatively simple structure. However, when aligning the TICs of chromatograms from complex mixtures with large concentration variability such as serum or urine, both algorithms often lead to misalignment of peaks and thus incorrect comparisons in the subsequent statistical analysis. This is mainly due to the fact that compounds with different m/z values but similar retention times are not considered separately but confounded in the benefit function of the algorithms using only one-dimensional information. Thus, it is necessary to treat the information of different mass traces separately in the warping function to ensure that compounds having the same m/z value and retention time are aligned to each other. The Component Detection Algorithm (CODA) is widely used to calculate the quality of an LC-MS mass trace. By combining CODA with the warping algorithms of DTW or PTW (DTW-CODA or PTW-CODA), we include only high quality mass traces measured by CODA in the benefit function. Our results show that using several CODA selected high quality mass traces in DTW-CODA and PTW-CODA significantly improves the alignment quality of three different, highly complex LC-MS data sets. Moreover, DTW-CODA leads to better preservation of peak shape as compared to the original DTW-TIC algorithm, which often suffers from a substantial peak shape distortion. Our results show that combination of CODA selected mass traces with different time alignment algorithm is a general principle that provide accurate alignment for highly complex samples with large concentration variability.

1 INTRODUCTION

Time alignment is a critical step in the data pre-processing of comparative studies based on LC-MS analyses, which are widely used in 'omics' experiments. Without an accurate time alignment, nonlinear retention time shifts across different chromatograms lead to incorrect peak matching and invalid subsequent statistical comparisons. This is especially the case for highly complex data sets with large concentration variation, where incorrect alignment may result in misinterpretation of comparative 'omics' experiments.[1-3] The importance of time alignment in comparative biomarker discovery studies is emphasized by the increasing number of published time alignment applications and review papers on the subject.[1, 4-29] These time alignment methods differ in their benefit functions as the criterion to construct the warping function that transforms the original retention time to the corrected retention time of the sample chromatogram. Since the throughput of proteomics experiments is constantly increasing, continuous improvement of automated time alignment methods is needed to align accurately the large amount of generated complex LC-MS data.

In this work, we focus on the modification of two widely used time alignment algorithms: Dynamic Time Warping (DTW)[5] and Parametric Time Warping (PTW)[6], and compare the results to our earlier work with the Correlation Optimized Warping-Component Detection Algorithm (COW-CODA).[4] Comparisons of time alignment quality of these algorithms using TICs or BPCs have been performed earlier for COW and DTW,[18, 25] COW and PTW,[26] or all of the three methods.[24] These comparisons were performed on simple data sets having a low number of compounds and low concentration variability. The studies concluded that generally COW improved peak alignment and resulted in close location of the corresponding peaks in different chromatograms, however in some cases, DTW resulted tighter alignment compare to COW. However, DTW is prone to distortion of peak shapes, while COW preserved well the peak shape after alignment. PTW was reported to be faster than DTW or COW but less precise in terms of time alignment.[24]

Several modifications have been introduced to these algorithms in order to improve the quality of time alignment and to adapt to different data sets. DTW has been adapted to align data sets derived from capillary electrophoretic,[24] gas chromatographic,[18, 25, 30] and near infrared spectroscopic data sets.[20] Peak shape distortions were observed and identified as major disadvantages of the DTW algorithm. This led to further work in order to preserve peak shape.[10, 25] Tomassi *et al.*[25] showed that changing the value of the slope constraint affected the extent of peak shape distortion after alignment. The optimum value was obtained empirically depending on the characteristics of the data (e.g. the initial retention time shift). Clifford *et al* introduced a variable penalty for each non-diagonal move in the warping path.[10] Even though these modifications were able to retain the peak-shape after DTW, they were intended to work on one-dimensional profiles only.

We show that both DTW and PTW fail in aligning the same compounds across multiple LC-MS data sets from complex proteomics or metabolomics samples, where many compounds with high concentration variability elute at similar retention times. The reason for this failure is that neither TICs nor BPCs provide information about m/z values for the benefit functions of DTW or PTW. Some approaches that take

the mass spectrometric information into account have been published. Most of these approaches do not work on the full data but rather on peak lists that require prior peak picking using various algorithms.[19, 21-23] The quality of time alignment using peak lists thus depends considerably on the quality of the peak detection algorithms.

Only a few approaches use single stage mass spectra and thus separate the intensity information for peaks of different masses eluting at the same retention time.[19, 31, 32] The first two methods either use the entire mass spectrum and a complex gap penalty functions to avoid a large number of consecutive non-diagonal steps in DTW,[19, 31] or use score functions which involve all peaks in the mass spectra and try to calculate the score due to the pure signal by removing noise contribution obtained with random reordering of peaks within mass spectra and setting each score below 0.2 to 0.[31] A third method use the 200 mass traces containing the highest peaks in the chromatograms.[32] The latter method provides insufficient description since it does not describe the form of the benefit function and the exact method used to combine the intensity information of different mass traces.

In the present studies, we are using a Component Detection Algorithm (CODA)[33] to select high-quality mass traces from a complete chromatogram prior to alignment. We subsequently separate the signals of the selected mass traces in the benefit function of DTW and PTW by summing up the differences between sample and reference chromatograms using each selected mass trace separately.

Combining CODA with DTW or PTW required fundamentally different mathematical approaches as compared to COW,[4] since the selection of high quality mass traces is performed prior to the alignment procedure while in COW-CODA mass trace selection is part of the warping procedure and different mass traces are selected for each COW segment. The alignment process of DTW and PTW, however, uses the same set of mass traces over the entire chromatographic time range and mass traces selection must be done prior to warping.

Performance of DTW-CODA and PTW-CODA was compared to each other as well as to COW-CODA and to the one-dimensional time alignment approaches (DTW-TIC, PTW-TIC and COW-TIC). The sum of overlapping peak areas was used as criterion to judge the time alignment quality on label-free single stage LC-MS data sets obtained during comparative profiling studies for biomarker discovery using trypsin-digested human serum and acid-precipitated urine samples.

2 MATERIAL AND METHODS

2.1 Computational Methods

Time alignment is driven by time concordance of common peaks (compounds) that are shared between reference and sample chromatograms. The success of such a procedure depends on the capacity of an algorithm to find as many common peaks as possible with high accuracy and to use only the information from these common peaks in the retention time shift correction procedure (benefit function). In general, a time alignment algorithm for LC-MS data must have the following properties: 1) assuring that peaks with similar retention times but different m/z values

are considered as separate features in the benefit function to avoid merging of different peaks signal having similar retention time but different m/z values; 2) considering only data from peaks that are shared between the reference and sample chromatograms in the benefit function; 3) discarding data containing a high level of noise; 4) taking peaks, background and noise distribution in the LC-MS data set into account locally; and 5) assuming that there are no changes in elution order of analytes between different LC-MS chromatogram.[8, 31]

Taking local peak distribution into consideration is more difficult using DTW or PTW than COW. This is because DTW performs retention time alignment data point-by-data point instead of the segment-wise procedure of COW. Selection of local high-quality mass traces, which are the same in the reference and sample chromatograms, but which could be different for each time point is not possible in DTW, because changing mass traces for different retention time data points will lead to discontinuity in the calculated minimal cumulative distance used in the benefit function of the algorithm. As is the case for DTW, it is not possible to use different mass traces at different retention times in PTW, since PTW computes the warping function using an iteration procedure. In each step it calculates the quadratic distance of the two traces using the entire time range, and locally different mass traces would result in similar discontinuity as with DTW. For that reason we have introduced a global mass trace selection procedure, to measure the quality of mass traces across the entire chromatogram based on the average local quality of the mass traces. We have further used these selected high quality mass traces for the entire retention time range in the warping procedure.

2.1.1 Measuring the average quality of LC-MS mass traces

This section describes how the local quality of a chromatogram is determined by measuring the quality of a mass trace and how high-quality mass traces are selected prior to the warping procedure. The quality of a chromatogram corresponds to the ratio between peak related information and noise. Three main types of noise in mass spectrometry data are spikes, chemical noise and electronic noise. Signals that correspond to a single data point, the so-called spikes, are generated at the ion source between the LC and MS and the MS ion optics interface.[34] The other two noise components are due to ionized, contaminating chemical compounds (chemical noise)[35] and to the electronic noise from the detector. The local average of the combined noise is the local background level of the chromatogram.

A mass trace with high background will have a high mean value, thus the mean subtracted mass trace will be rather different from the original signal. Similarly, a mass trace with spikes will differ strongly from the smoothed version that is obtained by using a moving average across several data points, as spikes are usually single-point events. Combining the measures of similarity between the mean-subtracted version and the smoothed version using a moving average of the original mass traces gives a single similarity value that takes both the contamination with spikes and the chemical and electronic background noise into account to result in a so-called quality index the Mass Chromatographic Quality (MCQ) after Windig *et al.*[33] High quality mass traces contain low noise levels and low spikes relative to the intensity of detected peaks. The CODA algorithm selects mass traces containing a large number of high intensity peaks by

calculating the MCQ of single mass chromatograms over the entire time range. However, in complex LC-MS data it is often the case that the quality of LC-MS signals varies with respect to retention time even within a single mass trace. For the DTW and PTW algorithm, it is preferable to select a mass chromatogram containing a large number of peaks more or less evenly distributed across the entire time range rather than mass traces with similar MCQ values but containing few peaks that are concentrated in a narrow retention time window. We have therefore modified CODA to take the local peak distribution into account giving preference to mass traces that contain an evenly distributed high number of intense peaks.

The quality of the local signal for each mass trace is measured by applying CODA to overlapping moving windows with a length of a data points, where a is an odd number so that integer b satisfies $a = 2b + 1$. Chromatogram $\boldsymbol{C}$ has size $m \times t$ where m corresponds to the index of mass traces and t to the index of retention times. For each position (i,j) in chromatogram $\boldsymbol{C}$, an MCQ value of $C(i,j-b...j+b)$ is calculated. This MCQ value is regarded as the quality of the signal at position (i,j). This step produces a matrix $\boldsymbol{Q}$, with the same size as the respective chromatogram, containing MCQ values for each data point of the chromatogram based on the equation below:

$$Q(i,j) = \begin{cases} CODA(C(i, j-b...j+b)) & \\ CODA(C(i,1...2b+1)) & for\ j \le b \\ CODA(C(i,t-2b...t)) & for\ j > (t-b) \end{cases} \qquad \text{Eq. 1.}$$

The quality of a mass trace is the average of the local MCQ values obtained for the same mass trace. Each chromatogram has thus a corresponding vector with size equal to the number of mass traces m. This vector contains the average MCQ values for the respective mass traces, and is used as quality scores to select mass traces prior to DTW and PTW (DTW-CODA, PTW-CODA).

2.1.2 Mass Trace Selection for Time Alignment

The alignment of two chromatograms requires the selection of several mass traces based on their respective quality in sample and reference chromatograms. Suppose chromatograms $\boldsymbol{C}_R$ and $\boldsymbol{C}_S$ have respectively a mean MCQ vector A_R and A_S, where index R and S refer to reference and sample chromatograms, respectively. The product $A_S \cdot A_R$ indicates the combined quality of mass traces in both chromatograms. Mass traces that result the highest product are then selected to be included in the warping function. In this paper, the product of average MCQ values obtained with moving windows will be referred to as the "Local Component Detection Algorithm", abbreviated as LCODA.

2.1.3 Dynamic Time Warping combined with LCODA-Selected Mass Traces (DTW-CODA)

The DTW algorithm using one-dimensional signals has been described previously in a number of publications.[1, 5, 10, 18-20, 24, 25] This section describes an extension of DTW algorithm using selected high quality mass traces based on the LCODA procedure. The concept of this method is illustrated schematically in Figure 1.

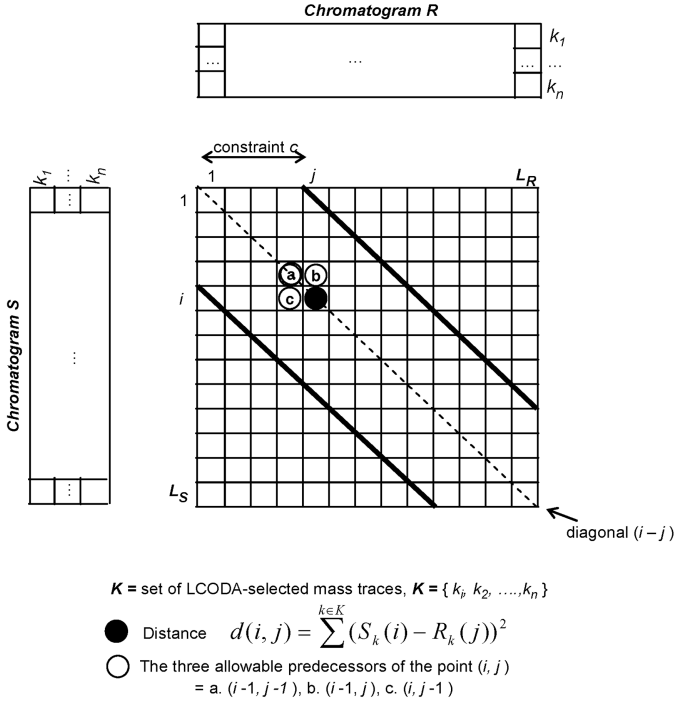

Figure 1. Schematic representation of the DTW-CODA algorithm. The grid representation shows the dynamic programming approach to calculate the cumulative minimal distance between sample *S* and reference *R* chromatograms using a set of LCODA-selected mass traces *K*. The area rounded by bold diagonal lines represents the search space of the DTW algorithm as defined by constraint *c*. In this area the cumulative minimum distance is calculated as the minimal sum of the intensities of LCODA-selected mass traces using the predecessor rules (see Equation 3) starting from grid point (1,1) until reaching the final grid point (L_S, L_R). Score and path matrices corresponding to the grid coordinates contain the cumulative minimum distance and the grid location indices of the points according to the cumulated minimal distance obtained using predecessor rules and the constraint *c*. The final optimal warping path is determined by backtracking the preceding grid indices consecutively starting from (L_S, L_R) until reaching (1,1).

Suppose two chromatograms ***R*** (reference chromatogram with number of time scans L_R) and ***S*** (sample chromatogram with number of time scans L_S) are to be aligned using a set of LCODA-selected mass traces ***K***, for each combination i and j, where $|i - j| \leq c$ (constraint). The local distance $d(i,j)$ is calculated by:

$$d(i,j) = \sum^{k \in K} (S_k(i) - R_k(j))^2 \qquad \text{Eq. 2.}$$

$\boldsymbol{S_k(i)}$ is the intensity value of mass trace k at time point i in chromatogram ***S*** and $\boldsymbol{R_k(j)}$ is the intensity value of mass trace k at time point j in chromatogram ***R***. Two matrices of size $L_S \times L_R$, which correspond to the grid presented in Figure 1, are constructed. The first matrix (*score matrix*) contains the minimal cumulative distances between ***S*** and ***R***. The second matrix (*path matrix*) contains the index of the optimum warping position that gives the respective cumulative minimum distance in the *score matrix*. The search space defining the allowed retention time transitions between the sample and reference chromatograms is defined by constraint c, which limits the maximal deviation from the diagonal by c number of points. For each position (i, j) in the defined search space, the minimum cumulative distances $D(i,j)$ are obtained from one of the three allowed predecessors $\{(i-1, j), (i-1, j-1), (i, j-1)\}$ based on Eq. 3. If the lowest score is obtained from different predecessors and one of them is the diagonal, then the diagonal path will be chosen to be included in the warping path.

$$D(i,j) = \min \begin{Bmatrix} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + d(i,j) \\ D(i,j-1) + d(i,j) \end{Bmatrix} \qquad \text{Eq. 3.}$$

The global minimal distance between chromatograms ***S*** and ***R*** is obtained by means of dynamic programming from the *score matrix* by calculating $D(i,j)$ from position (1,1) until (L_S, L_R) within the search space defined by constraint c. The global optimal warping path is obtained from the *path matrix* by backtracking the points resulting in the minimal cumulative distance as the last step of the time alignment procedure.

2.1.4 Parametric Time Warping combined with LCODA Selected Mass Traces (PTW-CODA)

The PTW algorithm for aligning one-dimensional signals between a sample $s(t_i)$ and a reference chromatogram $r(t_i)$ was introduced by Eilers.[6] The algorithm optimizes the coefficient a_d of the polynomial warping function $w(t_i)$ with d degrees so that the aligned sample signal $s(w(t_i))$ has the lowest cumulative distance G to the reference. In the present study we use a second-degree polynomial warping function with the form of $w(t_i) = a_0 + a_1 t_i + a_1 t_i^2$. The equation of the benefit function G is defined as follows:

$$G = \sum_{i \in H} [r(t_i) - \hat{s}(w(t_i))]^2 \qquad \text{Eq. 4.}$$

H indicates the set of indices i for which $\hat{s}(w(t_i))$ can be computed after interpolation of the sample chromatogram data points to the retention time vector (sampling points) of the reference chromatogram. In order to include the information of LCODA-selected traces, the benefit function G has been adapted to align two-dimensional signals based on a set of LCODA-selected traces $\boldsymbol{K}$ (Eq. 5). Based on a set of LCODA-selected traces one obtains one warping function $w(t)$ using iterative process, that is used to calculate a newly aligned retention time vector for the sample chromatogram.

$$G = \sum_{k \in K} \sum_{i \in H} [r_k(t_i) - \hat{s}_k(w(t_i))]^2 \qquad \text{Eq. 5.}$$

2.2 *Property of the Data Sets and Data Pre-Processing*

The algorithms were evaluated with three different LC-MS data sets with different analytical and biological characteristics as described in Christin *et al.*[4] Two data sets were derived from the analysis of trypsin-digested human serum (cervical cancer data set and factorial design data set) depleted of the six most abundant proteins and one data set from acid-precipitated urine of pregnant or non-pregnant women. The study protocol of the three data sets was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004. At the University Medical Center Groningen (UMCG, Groningen, The Netherlands) all newly referred patients are routinely asked to give written informed consent for collection and storage of pretreatment and follow-up serum, urine and tumor samples in a serum/urine/tissue bank for future research. Relevant patient data and follow-up are also retrieved and transferred into an anonymous, password-protected, database. According to Dutch regulations, these precautions mean no further institutional review board approval is needed (http://www.federa.org).

All chromatograms were acquired on an ion trap mass spectrometer (Agilent Technologies, LC-MSD SL series, Santa Clara, California, USA) in randomized order with automated gain control of accumulation time of ions to reach fixed number of 30.000 ions in the trap, using a rolling average of two spectra in single-stage MS mode. The acquired data were converted and subsequently stored in centroid mode.

2.2.1 Serum Samples

Serum samples were obtained from the Department of Gynecological Oncology (UMCG) and stored at -80°C in aliquots until analysis. Blood was collected in glass tubes (Becton Dickinson, Franklin Lakes, New Jersey, USA) with a siliconized inner wall, allowed to clot for at least 2 hours at room temperature before centrifugation at 1000 g for 10 minutes to obtain serum. Serum samples were stored at –80°C in the local serum bank until use. Before LC-MS analysis the serum samples were depleted of the six most abundant proteins using a Multiple Affinity Removal column (4.6 × 50 mm, Agilent Technologies). After trypsin digestion of the remaining proteins, all serum samples were stored at -80° C in aliquots until the final LC-MS analysis. The depletion protocol and trypsin digestion is described in detail in Govorukhina *et al.*[36]

2.2.1.1 *Cervical Cancer Data Set*

This data set is derived from a biomarker discovery study for cervical cancer. Serum samples from 10 patients were taken at two time points: before treatment (time point A) and after treatment with no recurrence of the disease for at least 6 months (time point B). These samples were analyzed by LC-MS resulting in 20 chromatograms. All patients in time point A showed high squamous cell carcinoma antigen-1 (SCCA-1) level (above 1.9 μg/ml) and no recurrence of disease after therapy except for two patients, one with partial remission and the other with stable disease where tumor remains without progression. The diagnosis was done by histological analysis and gynecological examination: inspection and palpation of the genitalia and SCCA-1 test. Patients with remission have no complains and normal SCCA-1 concentrations in time point B. All patients used in this study had advanced disease (stage III or IV) according to the International Federation of Gynecology and Obstetrics (FIGO) classification[37] and belonged to a group of long-term survivors. The level of the SCCA-1 was determined by ELISA.[38] Further details about the analysis of these samples using LC-MS are described in Govorukhina *et al.* [36]

2.2.1.2 *Factorial Design Data Set*

The serum sample for the factorial design study was obtained from one healthy female volunteer. The sample preparation procedure for this dataset was similar to the cervical cancer data set except for the following seven factors that were varied deliberately to investigate the influence of pre-analytical factors on the LC-MS profiles: blood collection tube, hemolysis level, clotting time, number of freeze-thaw cycles, trypsin to protein ratio, deactivation of trypsin after digestion, and stability of the digested sample in the autosampler of the LC-MS system at 4° C. Each factor was varied at two levels (high and low) and from 128 possible combinations, 16 combinations were selected according to a two-level 2_{IV}^{7-3} fractional design with resolution VI and with 3 repetitions of one condition. Detailed description of the factors and condition are described in Christin *et al.*[4] Nineteen LC-MS analyses were performed using the same protocol described by Govorukhina *et al.*[36]

2.2.2 Acid-precipitated Urine Data Set

Twenty-five first-void midstream morning urine samples from pregnant women were obtained from a local biobank (Department of Obstetrics and Gynecology of the University Medical Center in Groningen, The Netherlands) and directly stored frozen at -20° C. Twenty-five first-void midstream morning urine samples of non-pregnant women were collected in polypropylene containers and kept at 4° C for a maximum of 1 day before the samples were stored at -20° C in aliquots. The acid-precipitated urine samples were analyzed by LC-MS as described in Kemperman *et al.*[39]

2.3 Data Pre-processing

The original LC-MS chromatograms were converted to ascii files using the Bruker DataAnalysis (version 3.4, Build 181) software. Each ascii file was transformed into a two-dimensional matrix containing intensity values using an in-house developed data pre-processing pipeline. In this matrix each row has a corresponding m/z and each column a respective retention time value. During transformation, data reduction from 0.1 to 1 amu per bin was performed in the m/z dimension using two-dimensional Gaussian smoothing, while no data reduction was performed in the retention time dimension. All time alignment algorithms were applied to the transformed matrices. For each data set, one chromatogram was selected as the reference according to the procedure described in Christin *et al.*[4] Briefly, the best reference is the most frequently selected chromatogram having the highest sum of correlation to all other chromatograms in the data set based on the reconstructed TIC from a variable number of CODA selected mass traces. Similarly the worst reference is the most frequently selected chromatogram having the lowest sum of correlation to all other chromatograms in the data set of the reconstructed TIC from a variable number of CODA-selected mass traces. The list of the best and the worst reference for each data set can be found in the Table S-1 (Supporting Information).

A peak picking algorithm was applied to the transformed matrices based on a filter developed by Radulovic *et al.* called M-N rules[40] with $M = 8$ and $N = 3$. Signals are only retained if their intensity exceeds n times the local baseline for m consecutive data points in a single mass trace. The matrices obtained after peak picking are used later to evaluate and compare the quality of the time alignment algorithms by calculating the sum of overlapping peak area of two matrices. The data processing pipeline was executed on a personal computer equipped with an Intel® Core™ Quad CPU Q9300 @ 2.5 GHz processor and 8 GB of RAM. The time alignment software is written in Matlab and available at https://trac.nbic.nl/lcms_time_alignment_algorithms/.

3 RESULTS

3.1 Importance of data preprocessing

Ion trap mass spectrometers provide low resolution data, which contain small m/z shifts of peaks caused by local space charge effects.[41] Binning procedures summing up intensity in mass spectra between predefined borders are often used to reduce the amount of data and thus the processing time.[19] However, binning procedures processing centroided ion trap data result in very noisy data because of the local space charge effect. Application of two-dimensional smoothing using a Gaussian kernel, on the other hand, results in smooth data in both dimensions, which contain less noise, especially in the retention time dimension, which leads to a lower accumulated error in the benefit function of the time alignment algorithms (Figure S-1a,b). Figure S-1c shows an extracted ion chromatogram (EIC) of two adjacent masses of binned data and one mass trace of the data obtained after two-dimensional Gaussian smoothing of one peak. The binned data are fluctuating between the two adjacent mass traces, since the highest intensity is fluctuating between the borders of the bins. When mass spectra of binned

data are used to calculate the correlation, such fluctuations between two adjacent mass traces will result in noise, as the fluctuation is a random event with respect to different chromatograms. On the other hand, data obtained by two-dimensional Gaussian smoothing will result in smooth Gaussian type profiles for each of the m/z mass traces providing thus an efficient contribution to the correlation between two chromatograms. In our application we have used two-dimensional Gaussian smoothing to improve time alignment.

3.2 DTW-CODA and PTW-CODA

To compare the quality of the alignment with DTW and PTW in combination with CODA to the performance of the original algorithms (DTW-TIC and PTW-TIC respectively), we applied them to label-free LC-MS data from complex biological samples. To present the operating principle of the DTW/PTW CODA algorithms, two chromatograms from the urine data set were chosen randomly. We selected chromatograms from this dataset, because it is most challenging when it comes to time alignment problems due to the large concentration variation of compounds as a result of inter-individual (biological) differences.

3.2.1 Performance of DTW-CODA

An internal standard peptide (YPFPG, m/z 580, retention time 49.3 min), which was not included in the 200 selected mass traces, was used to assess the local time alignment quality. A visual comparison of the alignment of this peptide by DTW-CODA using different numbers of LCODA-selected traces shows that a minimum of 20 selected mass traces is needed to reduce the extensive misalignments that were observed with DTW-TIC. However, peak shape distortions were only avoided when the number of selected mass traces was extended to more than 50. Selecting 200 mass traces proved to give the best alignment with negligible peak shape distortion.

The difference in the optimal warping path obtained with DTW-TIC and DTW-CODA indicates that the algorithms do not arrive at the same final warping function (Figure 2). This is due to the fact that the TIC of the sample chromatogram is rather different from that of the reference chromatogram in the depicted region (45 – 55 min) making it difficult, if not impossible, for the DTW warping function to find the optimal warping path, since mass spectrometric information is not considered separately. The result is a 'random' warping path that leads to poor alignment and a distorted peak shape (Figure 3-middle panel). Simplifying the initial alignment problem by selecting 200 high-quality mass traces allows the DTW algorithm to find a warping path that deviates little from the diagonal of the *score* and *path matrices* reflecting the true small shifts between retention times in the original sample and reference chromatograms (see Figure 3-top panel). Such a "smoother" warping path leads to time alignment without peak distortion and tight alignment of peaks (see Figure 3-bottom panel). On the other hand, DTW-TIC will only succeed in aligning LC-MS data sets when the TIC profiles are "well-defined" and similar as a result of little concentration variability in the analyzed samples. For that reason, time alignment using the TIC of complex 'omics' LC-

MS analyses of body fluids containing many compounds with high concentration variability is challenging and in most cases impossible.

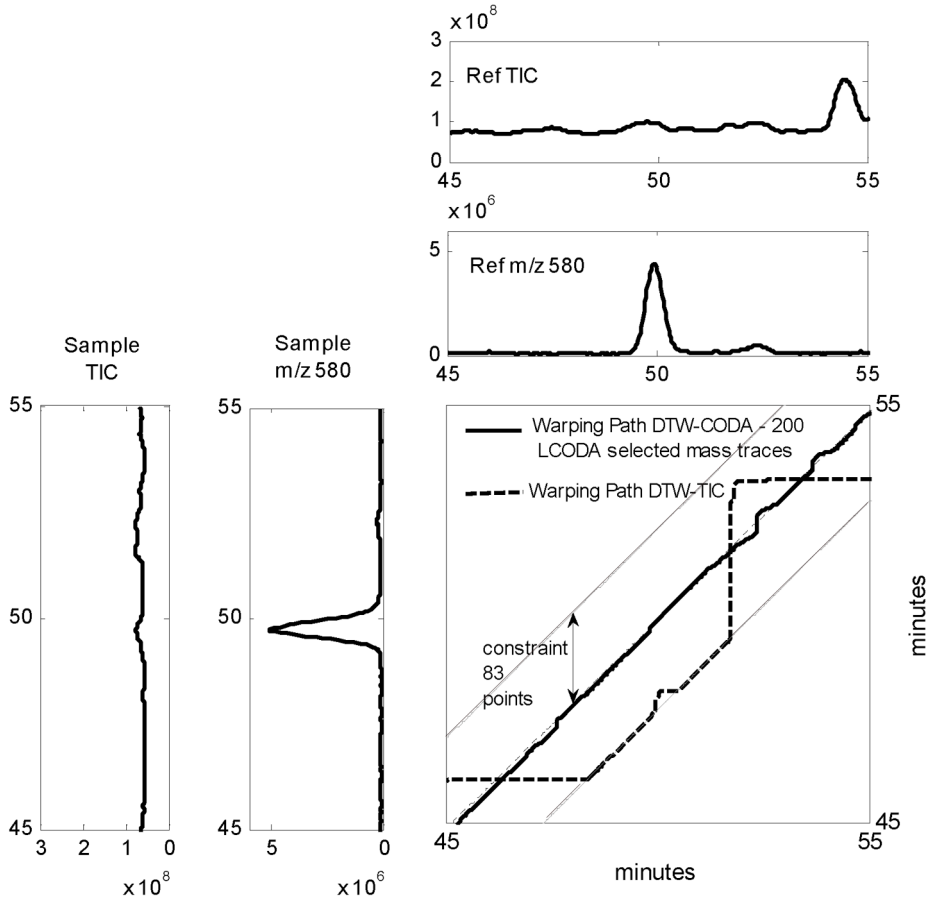

Figure 2. Comparison of the warping paths obtained with DTW-TIC and DTW-CODA using 200 high-quality LCODA-selected mass traces over a retention time window of 45-55 min in two chromatograms (5082628 and 5082630) from the acid-precipitated urine data set. DTW-TIC results in a rather 'chaotic' warping path ending in incorrect time alignment (Figure 3, middle panel). Using 200 pre-selected, high-quality mass traces, DTW-CODA follows a much smoother warping path correcting for the minor shifts in retention time that were present in the original data (Figure 3, top panel). Indeed DTW-CODA was able to find the optimal warping path by minimizing the sum of the cumulative distance of the LCODA-selected mass traces between the sample and reference chromatograms (see Figure 3, bottom panel).

Figure 3 shows the difference of the performance between DTW-TIC and DTW-CODA exemplified for a section of the EIC for m/z 580 related to the added standard peptide YPFPG. A major problem with the DTW-TIC algorithm is that it may lead to distortion of the chromatographic peak shape (see Figure 3-middle panel). When using several mass traces in the warping function of DTW-CODA, the algorithm tries to find the best compromise between alignments of each pair of peaks from different mass

traces. This results in a smoother warping path and a decreasing number of consecutive non-diagonal moves resulting in less peak distortion.

Figure S-2 (Supporting Information) shows the overlaid two-dimensional image of the reference and sample chromatogram obtained from urine samples with the original retention time, and after correcting retention time shifts using DTW-CODA (present the entire elution range of compounds (42-92 min) over an m/z range between 80-620 amu). This figure shows that non-linear retention time shifts in the sample chromatogram are accurately corrected with respect to the reference chromatogram by the DTW-CODA algorithm. A few peaks can be observed in the reference chromatogram (red) or sample chromatograms (green), which are absent in the corresponding other chromatogram (orphan peaks), while the shared common peaks present in both chromatograms (yellow) are well aligned. The retention times of these orphan peaks are also shifted by DTW-CODA to follow the same trend as the shared common peaks (e.g. peak at 335 m/z and 49 min of original retention time). This indicates that these peaks are also positioned correctly in the aligned chromatograms. In addition, the larger extent of retention time shifts observed in the beginning of the original chromatograms (20–40 min) (Figure S-2-b), most probably due to the use of a trapping column, were effectively corrected. One major advantage of combining DTW with LCODA-selected mass traces is that even without the use of special rules for the allowed predecessor steps in time alignment, the algorithm is highly conservative with respect to preserving the peak shape.

3.2.2 Performance of PTW-CODA

The same internal standard peptide (YPFPG, m/z 580, retention time 49.3 min) was used to assess the local alignment quality of PTW-CODA. To define the optimal number of selected mass traces, we calculated the sum of squared intensity differences using all points in single-stage LC-MS images between 8 randomly selected pairs of reference and the sample chromatograms after time alignment using a variable number of LCODA-selected mass traces ranging from 20 to 600. Two hundred selected mass traces resulted in the stabilized sum of squared intensity differences for these pairs of chromatograms (see Figure S-3 in Supporting Information). We have used the same 200 LCODA-selected mass traces for PTW-CODA as with DTW-CODA in order to facilitate comparison and to be sure that we are using a value, which is optimal for all chromatogram pairs.

The value of the 200 highest quality LCODA-selected mass traces gave accurate alignment in all three data sets. However, other data sets or different chromatogram pairs with different peak distribution or concentration variance may require a different number of high quality mass traces for optimal alignment. In that case the above described optimization procedure using the sum of squares of the intensity differences for all mass traces after applying the PTW-CODA algorithm should be used prior to the final application of PTW-CODA.

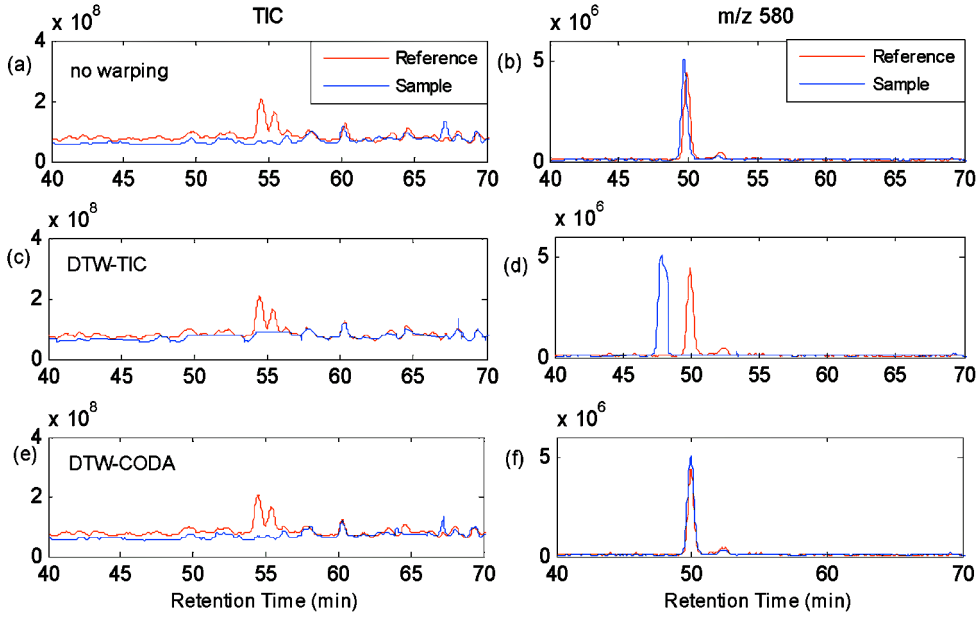

Figure 3. Application of DTW-TIC or DTW-CODA to two chromatograms obtained from acid precipitated urine sample (chromatograms 5082628 and 5082630; see also Figure 2. for the corresponding warping paths). a) original TICs (sample: blue; reference: red) prior to time alignment showing rather dissimilar profiles due to biological variability between samples; b) original EICs (580 +/- 0.5 amu) of the internal standard peptide YPFPG (m/z 580) prior to time alignment; c) TICs after time alignment with the DTW-TIC algorithm showing peak distortions at various locations; d) EICs (580 +/- 0.5 amu) of the internal standard peptide YPFPG after time alignment with the DTW-TIC algorithm showing major distortion of the peptide peak and a larger retention time shift compared to the original data; e) TICs after time alignment with the DTW-CODA algorithm showing tight alignment of the common peaks between sample and reference chromatograms even though the two most abundant peaks are absent in the sample chromatogram; f) EICs (580 +/- 0.5 amu) of the internal standard peptide YPFPG after time alignment with the DTW-CODA algorithm showing tightly aligned peaks without observable peak distortion.

The main parameter to choose for the PTW algorithm is the degree of the polynomial of the warping function. It has been investigated that the alignment quality using a cubic warping function did not result in a significant difference to the alignment quality using a quadratic warping function, since the coefficient for the highest degree term was always close to zero (results not shown). This means that the quadratic function was able to accurately adjust to the true form of the non-linear retention time shifts in the studied data sets. This may not be true for other data sets, where the true retention time shifts may have a more complex form. For that reason, if poor time alignment performance is observed with a quadratic function, using a higher order polynomial warping function may help to obtain a more accurate alignment. The only other user-defined parameter in PTW-CODA is the starting value of the coefficients of the warping function. As a starting point, we used the "no warping situation" in which the coefficients have the following values: $w(t) = t$, with $a_0 = 0$, $a_1 = 1$, and $a_2 = 0$.

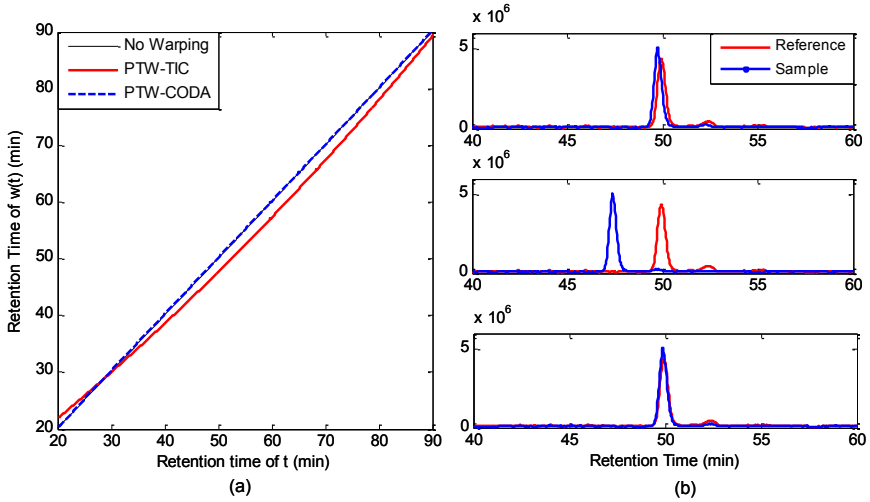

Figure 4. Comparison of the warping function of PTW-TIC and PTW-CODA using 200 high-quality LCODA-selected mass traces over a time window of 20-90 min applied to chromatograms 5082628 and 5082630 of the acid-precipitated urine data set. (a) shows the warping function obtained by PTW-TIC (red line) and the warping function obtained by PTW-CODA (blue line). The ellipse indicates the region presenting almost the largest retention time shifts of spiked standard peaks between the reference and sample chromatograms after applying PTW-TIC. (b) The EICs of a standard spiked peptide YPFPG (m/z 580 and retention time 49,34 min) before alignment (top panel), after alignment with PTW-TIC (middle panel) and after alignment with PTW-CODA using 200 LCODA-selected traces (bottom panel).

Figure 4 shows the comparison between PTW-TIC and PTW-CODA using 200 LCODA-selected mass traces. The warping function of PTW-TIC deviates strongly from the diagonal in comparison with the warping function of PTW-CODA, which resulted in major misalignment of the corresponding peaks (see Figure 4b, middle panel). The largest retention time shifts between identical peaks were observed in the middle of the chromatograms, where the retention time shifts after warping were actually larger than in the raw data. The ellipse in Figure 4a indicates the region with the largest retention time shift after alignment with PTW-TIC. This resulted major misalignment of the standard peptide YPFPG (EIC m/z 580 +/- 0.5 amu) (Figure 4b-middle panel). In contrast, PTW-CODA was able to correct the slight retention time shift between the original chromatograms of the standard peptide (Figure 4b, bottom panel).

3.3 *Comparison of DTW, PTW and COW coupled with LCODA-selected Mass Traces*

We have shown that the combination of DTW, PTW, and COW with CODA or LCODA significantly improves the alignment quality compared to the original algorithms that use the TIC in the benefit function. The question remains which time alignment algorithm combined with CODA or LCODA will provide the best time alignment for a given experimental LC-MS data set. In this section we compare the performance of DTW-CODA, PTW-CODA and COW-CODA by applying them to

experimental data sets with different compound distributions in m/z and retention time space (see Figure S-3 in Christin *et al*.[4]) and different concentration variability caused by contributions of various analytical and biological sources. The increasing order of the overall variance of the three data sets based on their respective relative standard deviation is: cervical cancer < factorial design < urine (see Figure 4c in Christin *et al*.).[4]

Concerning the COW-CODA algorithm, we have used the same values for the segment length and slack parameter as described previously.[4] Briefly, the segment length for the cervical cancer data set was 84 points (~1.5 min), for the factorial design data set 139 points (~2.3 min) and for the urine data set 83 points (~2.2 min). The slack parameter was set to 20% of the given segment length. In the COW-CODA algorithm, the selection of high quality mass traces by CODA was performed segment-wise with a maximum number of 30 traces per segment. In cervical cancer, factorial design and urine data sets, the number of selected mass traces during the time alignment procedure (union of all traces of all segments) was 507, 396, and 307, respectively. The value of the global constraint c in DTW was equal to the optimal segment length used in COW-CODA. Comparison of the performance of DTW-CODA and PTW-CODA was performed with 200 high-quality selected mass traces as discussed earlier. The constraint c has to be chosen to satisfy $c > |L_R - L_S|$, which is generally not a problem even for ion trap mass spectrometry data, where the data dependent accumulation time results in small difference in the number of data points. However, if large differences in sampling rate occur, such as warping a chromatogram acquired in single-stage MS to a chromatogram acquired in MS/MS mode, the 3-5 times differences in sampling rate of single-stage MS information should be corrected by interpolation in order to apply the DTW-based algorithm with success.

The large dynamic concentration range of analytes and the accurate quantification to detect discriminating compounds between healthy and diseased states requires that LC-MS data is acquired in single stage MS mode for biomarker discovery. In that case all time available for acquisition is used to collect quantitative information, while automated MS/MS would provide only quantitative information for every third or fifth scan time resulting in a decreased measured dynamic concentration range and less accurate quantification. After choosing the peaks of interest, the compound identity must be obtained from separate MS/MS measurements of pooled or individual samples.

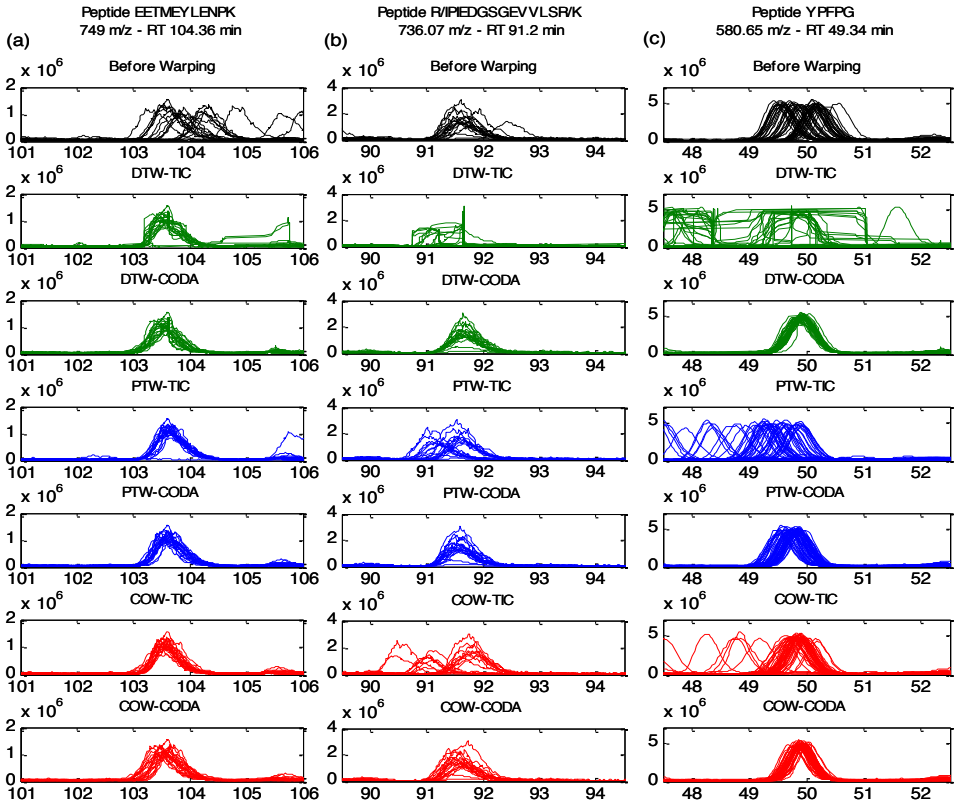

Figure 5. Local evaluation of the performance of the DTW-TIC and DTW-CODA (green) and the PTW-TIC and PTW-CODA (blue) algorithms in comparison with the previously described COW-TIC and COW-CODA (red) algorithms[1]. Extracted ion chromatograms show the retention time differences of spiked standard peptides in the cervical cancer (a), factorial design (b) and urine data sets (c) compared to the original data (top panel, black traces) using the best reference chromatograms. The *x*-axes correspond to the intensity and the *y*-axes correspond to the retention times (in minutes).

In order to assess the performance of the algorithms, we chose data sets acquired in single stage MS mode. However, it is difficult to assess the true performance of time alignment algorithms using single stage MS data, since peaks cannot be related to identified compounds as compared to data obtained with automated MS/MS data acquisition, where peak identity is generally used to assess the accuracy of time alignment[19] or the overall performance of time alignment/peak matching algorithms.[42] In consequence, the alignment quality was evaluated locally by comparing EICs of added internal standard peptides that were present in each chromatogram of a given data set before and after time alignment with DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA using the best reference chromatograms. Visualization of the EICs of one standard peptide each in the cervical cancer (a), factorial design (b) and urine (c) data sets served to judge the time alignment accuracy of the different algorithms visually (Figure 5). For all data sets the original DTW-TIC algorithm

showed the worst performance with considerable peak distortions and poor alignment while combining DTW with LCODA-selected mass traces (DTW-CODA) resolved these problems (see Figures S-4, S-5, S-6 in Supporting Information for other standard peptides). In general, all algorithms showed clearly improved alignment quality when combined with LCODA- or CODA-selected high-quality mass traces.

The initial concentration variability in the experimental LC-MS data sets plays an important role in the performance of the different time alignment algorithms. All algorithms (even DTW-TIC except for some remaining peak distortion) performed well on the cervical cancer data set (trypsin-digested serum; Figure 5a, left column) despite larger initial shifts in retention time, since the overall pattern was fairly conserved across the entire data set even at the TIC level. On the other hand, performance of time alignment methods was different for the factorial design data set (trypsin-digested serum; Figure 5b, middle column) containing larger analytical variability. For the factorial design data set all of the TIC-based algorithms did not resolved retention time shifts for the standard peptides and even increased retention time shifts with respect to the original data. Combination of all alignment algorithms with CODA- or LCODA-selected mass traces improved the overall alignment quality and resulted in tight peak clusters. A similar tendency was observed for the more variable urine data set, where the algorithms that work with CODA- or LCODA-selected mass traces improve alignment quality or at least maintain the original quality of the data, in cases where retention time shifts were already low (see also Figures S-4, S-5, and S-6 in Supporting Information).

Our earlier observation showed that time alignment with COW-CODA is not sensitive to the choice of reference chromatogram. We confirm this important behavior for DTW and PTW algorithms combined with LCODA-selected mass traces (Figures S-4, S-5, S-6, for alignment with the best and Figures S-7, S-8, S-9 for the worst reference in the Supporting Information). It is thus not necessary to select the optimal reference chromatogram. It is, however, noteworthy that large non-linear retention time shifts for peptides weakly binding to the chromatographic stationary phase, as sometimes observed at the beginning of the chromatographic elution gradient, are better corrected using the best reference chromatogram (see Figures S-6 and S-9 right column). The stability of the time alignment performance of algorithms using CODA- or LCODA-selected mass traces with respect to the reference chromatogram selection confirm further that all these methods results in tight alignment for the three experimental data sets.

To assess the global time alignment quality of different approaches, we compared the sum of the overlapping peak area between all possible pairs of chromatograms in the same data set as previously described.[4] An increased sum of the overlapping peak area is a measure of a globally improved time alignment. Figure 6 gives an overview of the sum of overlapping peak areas for each chromatogram with the remaining chromatograms in the three data sets using the best reference. The main observation is that all time alignment algorithms that make use of CODA- or LCODA-selected mass traces result in clearly increased overlapping peak areas when compared to the original data set independently of the variability in the original experimental data (Figure 6). The COW-TIC and the PTW-TIC algorithms show similar performance compared to the CODA-based algorithms for the cervical cancer data set, which contains

the lowest compound concentration variability (Figure 6a). This result is in agreement with the local evaluation using EICs of spiked standard peptides and can be explained with the well-defined character and high similarity of TIC traces of the chromatograms in this data set. It is noteworthy, that TIC-based algorithms may result in considerably higher retention time shifts than observed in the original data sets, especially for dataset with high compound concentration variability. For example, in the factorial design data set, which has the lowest initial retention time shifts, all algorithms using TICs resulted in lower overlapping peak area than the original chromatograms for the majority of samples (see Figure 6b). This effect was less pronounced in case of the urine data set, where alignment quality was largely unaffected by the TIC-based algorithms (see Figure 6c). The global assessment of DTW-TIC is more difficult because of the large extent of peak distortion. The observed lower overlapping peak area for all three data sets shows that this method is not appropriate to align complex LC-MS data sets. In contrast to DTW-TIC, the DTW approach combined with LCODA-selected mass traces resulted in vast improvements and makes the DTW-CODA algorithm the most accurate in many instances.

The sum of overlapping peak areas obtained using the worst reference resulted in similar results as with the best references (Figure S-10). This confirms that the algorithms using high-quality CODA- or LCODA-selected mass traces perform well independently of the reference chromatograms. This omits the application of a reference selection method in the time alignment procedure.

4 CONCLUSIONS

Complex LC-MS data sets with many overlapping peaks in the retention time dimension are difficult to compare unless one can assure that compounds are correctly matched prior to statistical analysis. We show that time alignment algorithms, that were originally developed for rather simple data sets, cannot be applied to this complex situation, as they either do not improve alignment or even make time alignment worse when using the TIC in a one-dimensional benefit function. We show furthermore that it is possible to simplify the initial complex data set by selecting m/z traces based on their respective Mass Quality Indices (MCQ values) using a modification of the CODA algorithm originally described by Windig.[33] The combination of DTW, PTW or COW algorithms with CODA-based trace selection, considering the different selected mass traces separately in the benefit function, resulted in clear time alignment improvements in three different complex LC-MS data sets containing increasing levels of concentration variability. Local and global assessment of the performance of the new algorithms showed that they were successful in aligning complex data sets as obtained during biomarker discovery and other quantitative comparative proteomics or metabolomics studies. Furthermore the time alignment algorithms using CODA- or LCODA-selected mass traces do not increase the number of parameters that need to be optimized. The optimal number of selected mass traces can be obtained from the data itself through optimization procedures.

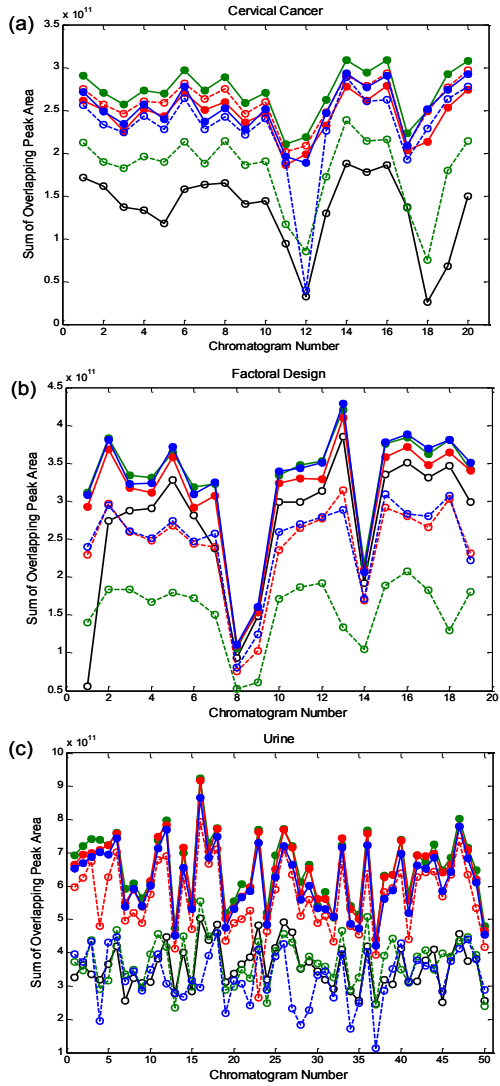

Figure 6. Sum of overlapping peak area of all chromatogram pairs using the best reference after applying M-N rules as peak filter to the cervical cancer (a), factorial design (b), and urine (c) data sets. The original chromatograms before time alignment (black) are compared to the chromatograms obtained after alignment with COW (red), PTW (blue) and DTW (green) using TICs (dashed lines, empty circles) or CODA-/LCODA-selected mass traces (full lines, filled circles). The chromatogram names corresponding to the chromatogram indices in the figures are reported in Supporting information (Table S2, S3, S4).

A distinct advantage of DTW-CODA algorithm is that it does not use any gap penalty function next to the constraint *c* to limit the search space. This facilitates the use of this algorithm; as compared to other versions of the DTW algorithm using separate mass information in their benefit functions.[19, 31] The form of the benefit function, may also play a role. Future research should focus on better optimization of the form of the

benefit function, such as using the cumulative covariance or the cumulative correlation.[19] We did not explore these possibilities, since we obtained highly accurate time alignments for all pairs of chromatograms in the studied data sets withthe cumulative sum of quadratic distances as benefit function for DTW or PTW. We have shown that combination of CODA-selected mass traces with different time alignment methods is a general principle to align complex LC-MS data sets with high compound concentration variability. The main characteristics of the three time alignment algorithms based on our implementation in this work are presented in Table 1.

Although all three time alignment algorithms using mass trace selection perform similarly well on highly complex LC-MS data sets, there are certain features, which discriminate them from a user perspective. The need to set parameters may complicate the proper use of an algorithm. User-defined values for parameters are an important point for some time alignment algorithms, since they affect the results significantly and should be adapted to the characteristics of the data sets (e.g. initial retention time shifts, average peak width, concentration variability). PTW-CODA has a distinct advantage in this respect for it does not require the user to set any parameters prior to starting the alignment procedure. The degree of the polynomial order of the warping function and the initial setup of the polynomial coefficients does not affect the alignment result, which is robust with respect to the changes of key analytical properties of the datasets such as peak distribution or concentration variation. Second, the requirements for computing capacity may be a limiting factor for users, especially if they do not have access to distributed computing facilities. DTW-CODA is advantageous in this respect, as it completes one round of time alignment for two complex LC-MS data sets in about 15 sec on a powerful personal computer (Intel® Core™ Quad CPU Q9300@2.5 GHz processor and 8 GB of RAM) while COW-CODA takes about 12 min. The significantly different execution time of COW-CODA is due to the fact that this algorithm includes segment-wise CODA trace selection as part of the warping function, while both DTW- and PTW-CODA require prior selection of high-quality mass traces, which takes 12 min per chromatogram. However, once the quality measurement by the LCODA procedure has been performed, this LCODA matrix can be reused for both DTW-CODA and PTW-CODA making this a one-time investment in computing time per data set. A note of caution has to be added with respect to using the DTW-CODA algorithm, as it may still introduce peak distortions, albeit much less than the original DTW-TIC algorithm. This must be carefully evaluated and can be considered as the main disadvantage of this approach. However, if peak quantification is performed before the time alignment and the alignment results are only used for peak matching, small peak distortions do not affect the statistical outcome of the comparative profiling study. In this case time alignment will only affect peak clustering performance, which will be highly accurate as the chromatographic signal is tightly aligned in the dataset, even though the peaks are slightly distorted.

To compare the performance of time alignment algorithms, the global evaluation based on overlapping peak areas is a reliable guide. However, since minor peak distortions and local misalignments may still occur, it is recommended to inspect the aligned chromatograms also visually using EICs of defined peaks (e.g. added internal standards or CODA-selected traces) before and after alignment.

Characteristics	**DTW-CODA**	**PTW-CODA**	**COW-CODA**
Setting of parameters	One parameter: constraint *c* (in time) to limit the borders of the search space.	No user-defined parameters required.	Two parameters: segment length (in time) and slack parameter (in % of segment length).
Peak shape distortion	Minor	No	No
CODA trace selection method	Global trace selection using the LCODA procedure.	Global trace selection using the LCODA procedure.	Local, segment-wise trace selection during algorithm execution.
Execution time for one pair of chromatograms[1]	~15 sec excluding mass trace selection (12 minutes per chromatogram).	~ 1 min excluding mass trace selection (12 minutes per chromatogram).	12 minutes including mass trace selection.

Table 1. Main characteristics of DTW-CODA, PTW-CODA and COW-CODA.

[1] 7000 time scans, 200 selected mass traces, using Intel® Core™ Quad CPU Q9300 @ 2.5 GHz processor and 8 GB of RAM.

5 REFERENCES

1. Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J Chromatogr A* **2002,** 961, (2), 237-244.
2. Christensen, J. H.; Hansen, A. B.; Karlson, U.; Mortensen, J.; Andersen, O., Multivariate statistical methods for evaluating biodegradation of mineral oil. *J Chromatogr A* **2005,** 1090, (1-2), 133-45.
3. Bahowick, T. J.; Synovec, R. E., Sequential chromatogram ratio technique: evaluation of the effects of retention time precision, adsorption isotherm linearity, and detector linearity on qualitative and quantitative analysis. *Analytical Chemistry* **1992,** 64, (5), 489-496.
4. Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L., Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal Chem* **2008,** 80, (18), 7012-7021.
5. Kassidas, A.; MacGregor, J. F.; Taylor, P. A., Synchronization of batch trajectories using dynamic time warping. *AIChe* **1998,** 44, 864.
6. Eilers, P. H. C., Parametric time warping. *Anal Chem* **2004,** 76, (2), 404-411.
7. Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. r., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J Chromatogr A* **1998,** 805, 17-35.
8. Aberg, K. M.; Alm, E.; Torgrip, R. J., The correspondence problem for metabonomics datasets. *Anal Bioanal Chem* **2009,** 394, (1), 151-62.
9. Chae, M.; Reis, R. J. S.; Thaden, J. J., An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas chromatography-mass spectrometry peaks. *BMC Bioinformatics* **2008,** 9 Suppl 9, S15.
10. Clifford, D.; Stone, G.; Montoliu, I.; Rezzi, S.; Martin, F. o.-P.; Guy, P.; Bruce, S.; Kochhar, S., Alignment Using Variable Penalty Dynamic Time Warping. *Anal Chem* **2009**.
11. Finney, G. L.; Blackler, A. R.; Hoopmann, M. R.; Canterbury, J. D.; Wu, C. C.; MacCoss, M. J., Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution muLC-MS data. *Anal Chem* **2008,** 80, (4), 961-971.
12. Fischer, B.; Grossmann, J.; Roth, V.; Gruissem, W.; Baginsky, S.; Buhmann, J. M., Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* **2006,** 22 (14), 132-140.
13. Fischer, B.; Roth, V.; Buhmann, J. M., Time-series alignment by non-negative multiple generalized canonical correlation analysis. *BMC Bioinformatics* **2007,** 8 Suppl 10, S4.
14. Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R., Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. *J Am Soc Mass Spectrom* **2007,** 18, (10), 1835-1843.
15. Paulus, C.; Bonnet, S.; Gerfault, L.; Mery, E.; Strubel, G.; Ricoul, F.; Grangeat, P., Chromatographic alignment combined with chemometrics profile reconstruction approaches applied to LC-MS data. *Conf Proc IEEE Eng Med Biol Soc* **2007,** 2007, 5984-5987.
16. Pierce, K. M.; Wood, L. F.; Wright, B. W.; Synovec, R. E., A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data. *Anal Chem* **2005,** 77, (23), 7735-7743.
17. Pierce, K. M.; Wright, B. W.; Synovec, R. E., Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatographic data using the piecewise alignment algorithm. *J Chromatogr A* **2007,** 1141, (1), 106-116.
18. Pravdova, V.; Walczak, B.; Massart, D. L., A comparison of two algorithms for warping of analytical signals. *Anal Chim Acta* **2002,** 456, 77-92.
19. Prince, J. T.; Marcotte, E. M., Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Anal Chem* **2006,** 78, (17), 6140-6152.
20. Ramaker, H.-J.; van Sprang, E. N. M.; Westerhuis, J. A.; Smilde, A. K., Dynamic time warping of spectroscopic BATCH data. *Anal Chim Acta* **2003,** 498, 133-153.

21. Sadygov, R. G.; Maroto, F. M.; Huhmer, A. F., ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Anal Chem* **2006,** 78, (24), 8207-17.
22. Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O., OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008,** 9, 163.
23. Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P., Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal Chem* **2008,** 80, (9), 3095-3104.
24. Szymaska, E.; Markuszewski, M. J.; Capron, X.; van Nederkassel, A.-M.; Heyden, Y. V.; Markuszewski, M.; Krajka, K.; Kaliszan, R., Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides. *Electrophoresis* **2007,** 28, (16), 2861-2873.
25. Tomasi, G.; van den Berg, F.; Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemometrics* **2004,** 18, 231-241.
26. van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H. C.; Heyden, Y. V., A comparison of three algorithms for chromatograms alignment. *J Chromatogr A* **2006,** 1118, (2), 199-210.
27. van Nederkassel, A. M.; Xu, C. J.; Lancelin, P.; Sarraf, M.; Mackenzie, D. A.; Walton, N. J.; Bensaid, F.; Lees, M.; Martin, G. J.; Desmurs, J. R.; Massart, D. L.; Smeyers-Verbeke, J.; Heyden, Y. V., Chemometric treatment of vanillin fingerprint chromatograms. Effect of different signal alignments on principal component analysis plots. *J Chromatogr A* **2006,** 1120, (1-2), 291-298.
28. Wang, P.; Tang, H.; Fitzgibbon, M. P.; McIntosh, M.; Coram, M.; Zhang, H.; Yi, E.; Aebersold, R., A statistical method for chromatographic alignment of LC-MS data. *Biostatistics* **2007,** 8, (2), 357-367.
29. Zhang, D.; Huang, X.; Regnier, F. E.; Zhang, M., Two-dimensional correlation optimized warping algorithm for aligning GC x GC-MS data. *Anal Chem* **2008,** 80, (8), 2664-2671.
30. Vial, J.; Nocairi, H.; Sassiat, P.; Mallipatu, S.; Cognon, G.; Thiebaut, D.; Teillet, B.; Rutledge, D. N., Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. *J Chromatogr A* **2009,** 1216, (14), 2866-72.
31. Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B., Signal Maps for Mass Spectrometry-based Comparative Proteomics. *Mol Cell Proteomics* **2006,** 5, (3), 423-432.
32. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **2003,** 75, (18), 4818-26.
33. Windig, W.; Phal, J. M.; Payne, A. W., A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Anal Chem* **1996,** 68, 3602--3606.
34. Windig, W.; Smith, W. F.; Nichols, W. F., Fast interpretation of complex LC/MS data using chemometrics. *Analytica Chimica Acta* **2001,** 446, (1-2), 465-474.
35. Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M., Interferences and contaminants encountered in modern mass spectrometry. *Anal Chim Acta* **2008,** 627, (1), 71-81.
36. Govorukhina, N. I.; Reijmers, T. H.; Nyangoma, S. O.; van der Zee, A. G. J.; Jansen, R. C.; Bischoff, R., Analysis of human serum by liquid chromatography-mass spectrometry: improved sample preparation and data analysis. *J Chromatogr A* **2006,** 1120, (1-2), 142-150.
37. Benedet, J. L.; Bender, H.; Jones, H., 3rd; Ngan, H. Y.; Pecorelli, S., FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers. FIGO Committee on Gynecologic Oncology. *Int J Gynaecol Obstet* **2000,** 70, (2), 209-62.
38. Esajas, M. D.; Duk, J. M.; de Bruijn, H. W.; Aalders, J. G.; Willemse, P. H.; Sluiter, W.; Pras, B.; ten Hoor, K.; Hollema, H.; van der Zee, A. G., Clinical value of routine serum squamous cell carcinoma antigen in follow-up of patients with early-stage cervical cancer. *J Clin Oncol* **2001,** 19, (19), 3960-6.
39. Kemperman, R. F. J.; Horvatovich, P. L.; Hoekman, B.; Reijmers, T. H.; Muskiet, F. A. J.; Bischoff, R., Comparative urine analysis by liquid chromatography-mass spectrometry and

multivariate statistics: method development, evaluation, and application to proteinuria. *J Proteome Res* **2007,** 6, (1), 194-206.

40. Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A., Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry. *Mol Cell Proteomics* **2004,** 3, (10), 984-97.
41. Cox, K. A.; Cleven, C. D.; Cooks, R. G., Mass shifts and local space charge effects observed in the quadrupole ion trap at higher resolution. *International Journal of Mass Spectrometry and Ion Processes* **1995,** 144, (1-2), 47-65.
42. Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C., Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **2008,** 9, (1), 375.

Chapter 4

A critical assessment of statistical methods for biomarker discovery in clinical proteomics

Submitted

ABSTRACT

In this chapter, we compare the performance of six different feature selection methods for LC-MS based proteomics and metabolomics biomarker discovery: *t*-test, Mann-Whitney-Wilcoxon-test (*mww*-test), Nearest Shrunken Centroid (NSC), linear Support Vector Machine – Recursive Features Elimination (SVM-RFE), Principal Component Discriminant Analysis (PCDA) and Partial Least Squares Discriminant Analysis (PLSDA) using urine samples that were spiked with a range of peptides at different concentration levels. The ideal feature selection method should select the complete list of discriminating features that are related to the spiked peptides. While many studies have to rely on classification error to judge the reliability of the selected biomarker candidates, we assessed the accuracy of selection directly from the list of spiked peptides. The statistical methods were applied on data sets with different sample size and extent of sample class separation determined by the concentration level of spiked compounds. For each statistical method and data set, the performance for selecting a set of features related to spiked compounds was assessed using the harmonic mean of the recall and the precision (*f*-score) and the geometric mean of the recall and the true negative rate (*g*-score). We conclude that the univariate t-test and the *mww*-test with multiple testing correction are not applicable to data sets with small sample size (n=6), but that their performance improves markedly with increasing sample size. PCDA and PLSDA select small feature sets with high precision, NSC strikes a reasonable compromise between recall and precision for all data sets while linear SVM-RFE performs poorly for selecting features related to the spiked compounds.

1 INTRODUCTION

Biomarker discovery plays an important role to advance medical research by allowing early diagnosis of disease and prognosis of treatment interventions [1, 2]. Biomarkers may be proteins, peptides or metabolites as well as mRNAs or other kinds of nucleic acids (*e.g.* microRNAs) whose level changes in relation to the stage of a given disease and that may further be used to accurately assign the disease stage of a patient based on a biochemical test. The accurate selection of biomarker candidates is crucial, since it determines the outcome of further validation studies and the ultimate success of developing diagnostic and prognostic assays with high specificity and sensitivity. The success of biomarker discovery depends on several factors: consistent and reproducible phenotyping of the individuals from whom biological samples are obtained, the quality of the analytical methodology which in turn determines the quality of the collected data, the accuracy of the computational methods to extract quantitative information to define the molecular identity of biomarker candidates from raw analytical data and finally the applied statistical method to select a limited list of compounds with the potential to discriminate between predefined classes of samples. *De novo* biomarker research consists of a biomarker discovery and a biomarker validation part [3]. Biomarker discovery uses analytical techniques, which try to measure as many compounds as possible in a relatively low number of samples. The goal of subsequent data preprocessing and statistical methods is to select a limited number of compounds, which are subsequently subjected to targeted analysis in large number of samples. Advanced technology, such as high-throughput and high-resolution liquid chromatography – mass spectrometry (LC-MS), is increasingly applied in biomarker discovery research. Such analyses detect tens of thousands of compound and background-related signals in a single biological sample generating enormous amounts of multivariate data. Data preprocessing workflows reduce data complexity considerably by extracting only the information related to compounds resulting a quantitative feature matrix, where rows and columns correspond to samples and extracted features or *vice versa*. Features are data preprocessing components, which are mainly related to sample-constituting compound peaks, such as isotopologues for data obtained with high resolution mass spectrometers or peaks corresponding to the average mass in data acquired with low resolution mass spectrometers. Features may also be related to data preprocessing artifacts, and the ratio of such erroneous features to peptide-related features depends on the performance of the data preprocessing workflow. Preprocessed LC-MS data sets contain a large number of features compared to sample size. These features are characterized by their m/z value and retention time, and in the ideal case they are combined and linked to compound identity such as metabolites, peptides and proteins. In LC-MS based proteomics and metabolomics studies, sample analysis is time consuming so that it is impossible to increase the number of samples to a level to balance the number of features in a data set. Therefore the success of biomarker discovery depends on powerful biomarker candidate selection methods that can deal with a low sample size and a high number of features. Due to the unfavorable statistical situation, it is, however, pivotal to validate the discovered biomarker candidates in a set of independent samples, preferably in a double-blinded fashion [1].

Biomarker selection is often based on classification methods, which are preceded by feature selection methods (filters), or have built-in techniques (wrappers and embedded) to select a list of compounds/peaks/features, which provide the best classification performance on predefined sample groups (e.g. healthy and disease) [4]. Classification methods with feature selection can also be used as classifiers (e.g. to classify an unknown sample into a predefined sample class), however they should be distinguished from classifiers having no feature selection option, but perform the classification using all variables. A few feature selection methods such as filters can be used independently of the classifiers. Classification methods without feature selection ability cannot be used for biomarker discovery because these methods aim to classify samples into predefined classes correctly but cannot identify the variables (features/compounds) causing the separation of sample groups [5, 6].

Different statistical methods with feature selection have been developed according to the complexity of the analyzed data, and have been extensively reviewed [4, 5, 7, 8]. Ways of optimizing such methods, to improve sensitivity and specificity, have become a major topic in biomarker discovery research and in the many 'omics-related' research areas [5, 9, 10]. Some comparisons of classification methods related to their classification and learning performance have been initiated. Dougherty *et al.* [11] studied the properties of feature selection methods, such as the relation of the selected feature sets to the theoretically best feature sets based on the classification error and optimal number of selected features, observation of the 'peaking phenomenon' and the accuracy of feature selection methods in high-dimensional data. Another study focused on finding the most accurate classifiers for simulated data sets with sample sizes ranging from 20 to 100 [12]. Rubingh *et al.* [13] compared the influence of sample size in an LC-MS metabolomics data set on the performance of three different statistical validation tools: cross validation, jack-knifing model parameters, and a permutation test. This study concluded that for small sample sets the outcome of these validation methods is influenced strongly by individual samples and therefore cannot be trusted and the validation tool cannot be used as warning mechanism for problems due to sample size or representability of sampling. This implies that reducing the dimensionality of the feature space is critical when approaching a classification problem where the number of features exceeds the number of samples by a large margin. Dimensionality reduction, whether it is built into (such as wrappers and embedded methods) or independent of a classification method (filters), retains a smaller set of features to bring the feature space in line with sample size to allow application of statistical methods, which perform only with acceptable accuracy when sample and feature size are similar.

In this study we compared different classification methods focusing on feature selection in a spiked LC-MS data set that mimics the situation of a biomarker study. Our results provide guidelines to researchers that will engage in biomarker discovery or other differential profiling 'omics' studies with respect to sample size and to selecting the most appropriate classification method with feature selection for a given data set. We evaluated the following approaches: univariate *t*-test, Wilcoxon-Mann-Whitney test (*mww*-test) with multiple testing correction [14], Nearest Shrunken Centroid (NSC) [15, 16], Support Vector Machine – Recursive Features Elimination (SVM-RFE) [17], Partial Least Squares – Discriminant Analysis (PLSDA) [18], and Principal Component – Discriminant Analysis (PCDA) [19]. PCDA and PLSDA were combined with the rank-

product as feature selection criterion [20]. These methods were evaluated with data sets having two characteristics: varying sample size and varying class separation due to different concentration levels of the added compounds. Data were acquired by LC-MS from urine samples that were spiked with a set of known peptides (true positives) at different concentration levels. These samples were then combined in two classes containing peptides spiked at low and high concentration levels. The performance of the classification methods with feature selection was measured by their ability to select features that are related to the spiked peptides. Since true positives are known in our data set, we quantify performance by the *f*-score (the harmonic mean of precision and recall) and the *g*-score (the geometric mean of accuracy).

2 EXPERIMENTAL PROCEDURES

2.1 Dataset Design

Fifty urine samples were obtained from 15 healthy females and 35 healthy males over the age range of 26.9 to 72.9 years. Two hundred μL were taken from each sample creating one pooled urine sample. The pooled urine sample was spiked with a tryptic digest (Promega, Madison, WI, V5111) of carbonic anhydrase (Sigma, Steinheim, Germany, C3934) as well as with seven synthetic peptides at 8 different dilutions: 6.25, 12.5, 25, 50, 100, 200, 400, and 2000 times dilution (called groups A-H, respectively), of the stock solution containing 240 μM trypsin-digested carbonic anhydrase and the following concentrations (in μM) for the seven synthetic peptides: VYV, 83; YGGFL, 57; DRVYIHPF, 29; YPFPGPI, 46; YPFPG, 60; GYYPT, 54; and YGGWL, 57. At each concentration level, the sample was analyzed 5 times resulting in 40 LC-MS chromatograms. These chromatograms were pre-processed with constant resolution of 0.1 amu [21, 22] resulting in a final common peak list of 29529 peaks with 151 peaks originating from the added peptides. Details on sample preparation and LC-MS data acquisition is provided in Supporting Information.

From these 40 chromatograms, six data sets consisting of two classes with 2 different levels of class separation and 3 different sample sizes were designed following the scheme in Table 1. These six datasets were combined into Data set 1 (large class separation) and Dataset 2 (small class separation) and 3 different sample sizes (indicated with a, b and c; see Table 1 for details). Data set 1 contains samples from groups A-C as high concentration spiked class (class 1) and from groups F-H as low concentration spiked class (class 0). Data set 2 contains groups B-D as high concentration spiked class (class 1) and groups E-G as low concentration spiked class (class 0). From each data set, three different subsets were formed with 6, 12 and 15 samples per class, respectively. For data sets with a total of 6 samples per class, 2 samples were taken randomly from each group (A-H). For data sets with a total of 12 samples per class, 4 samples were chosen randomly from each group. For data sets with a total of 15 samples per class, all samples were included. Feature selection for each combination of methods and datasets was repeated 100 times using each time a different combination of samples that were selected using the sample selection scheme described in the Table 1.

Data Set	Class Separation	Sample size per class	
Data Set 1a-c	Large class separation High spiked class = combination of groups A-C Low spike class = combination of groups F-H	1a	6: two samples were randomly taken from each of the groups A-C (class 1) and F-H (class 0)
		1b	12: four samples were randomly taken from each of the groups A-C (class 1) and F-H (class 0)
		1c	15: all samples from groups A-C (class 1) and F-H (class 0)
Data Set 2a-c	Small class separation High spiked class = combination of groups B-D Low spike class = combination of groups E-F	2a	6: two samples were randomly taken from each of the groups B-D (class 1) and E-F (class 0)
		2b	12: four samples were randomly taken from each of the groups B-D (class 1) and E-F (class 0)
		2c	15: all samples from groups B-D (class 1) and E-F (class 0)

Table 1. Description of the sample groups that were combined to give data sets 1a-c and 2a-c. This scheme was used to select files for the 100 repetitions of each combination of feature selection methods and data sets (see Table S-1 for results).

2.2 Biomarker discovery methods

2.2.1 Univariate Tests

The parametric univariate *t*-test ranks features according to their *p*-value and is not a classification method. Since the data sets contained six to fifteen samples per class, it is difficult to test the normality of the data, which in this case is the distribution of the peak intensities. We therefore also used a non-parametric univariate filter, the Wilcoxon-Mann-Whitney test (*mww*-test). Since the data sets contain a large number of features, we corrected the calculated *p*-values using the Benjamini-Hochberg approach [14]. A feature was considered significant when the *p*-value was below 0.05 after multiple testing correction.

2.2.2 Semi-Multivariate - Nearest Shrunken Centroid

The Nearest Shrunken Centroid (NSC) approach aims to find a set of features that gives the minimum classification error or the highest sum of correct class probability in a set of training samples using double cross-validation by progressively eliminating features that do not contribute to construction of the shrunken class centroid. This method was proposed by Tibshirani *et al.* for classification of cancer samples based on a microarray data set [15, 16]. The double cross-validation scheme for this method is outlined in the Figure 1. Other classification methods with feature selection used in this article were implemented according to similar double cross-validation schemes (see Figures S-1 and S-2, Supporting Information).

The distance d_{ik} between a feature *i* in class *k* and its respective overall centroid is calculated as the difference between the within class mean $\bar{x}_{ik}$ and the overall mean $\bar{x}_i$, standardized by the standard error. The standard error (Eq. 1) is calculated using the pooled within class standard deviation of the respective feature s_i, a constant s_0 (median of the standard deviation s_i across all features) to avoid large distances due to small standard deviations and the constant $m_k = \sqrt{1/n_k - 1/n}$. The shrinkage threshold Δ is iteratively subtracted from this distance, and features whose shrunken distance in all classes is zero or negative are eliminated. A test sample *x** is attributed to the class to which it has the highest class probability δ_k. The discriminant score for class *k* and for test sample *x** is $\delta_k(x^*)$, which is the sum of the standardized squared distances between each relevant feature in the test sample *x** and the k^{th} shrunken centroid $\bar{x}'_{ik}$ corrected by the prior probability π_k of class *k* (Eq. 2). This distance is basically similar to a simple diagonal covariance matrix between the test sample and the shrunken centroid of the respective class. Since feature elimination is done univariately, but classification of the test sample to the class-specific shrunken centroid at a given shrinkage is calculated multivariately, we call this a semi-multivariate method.

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)} \qquad \text{Eq. 1}$$

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log \pi_k \qquad \text{Eq. 2}$$

Based on the discriminating score, we can calculate the class probability $\hat{p}_k(x^*)$ (Eq. 3), which is the probability of sample *x** belonging to class *k*.

$$\hat{p}_k(x^*) = \frac{e^{-\frac{1}{2}\delta_k(x^*)}}{\sum_{l=1}^{K} e^{-\frac{1}{2}\delta_l(x^*)}} \qquad \text{Eq. 3}$$

Once the class probability has been calculated for each test sample, and for each shrinkage value, the probability of the true class for each sample is summed up. We now have two measurements based on which we select the optimum subset of features: a) the subset that minimizes the classification error (Eq. 2), and b) the subset that maximizes the sum of true class probabilities (Eq. 3). In our study the optimum shrinkage is chosen based on the maximum true class probability of the test data set, since it gives a continuous plot and a well-defined optimal shrinkage value. Once the optimum shrinkage has been obtained in the inner cross-validation loop (see Figure 1), it is applied to the training data set in the inner loop to obtain the optimal corresponding feature set. Based on this feature set, the discriminant and the true probability scores are calculated for the independent test data set in the outer loop to assess the classification model performance. Since each passage through the outer loop yields a different value for the optimum shrinkage, we calculate the median of the correct class probability

scores at each shrinkage value from all the outer loops at the end of the double cross validation scheme. Fluctuation of the true class probability curves as a function of the optimal shrinkage values obtained for various outer loop evaluations reflects the stability of the model and small class differences in the studied data set. The shrinkage at the maximum of the median true class probability curve is used to select the optimal feature set using all samples in the data set.

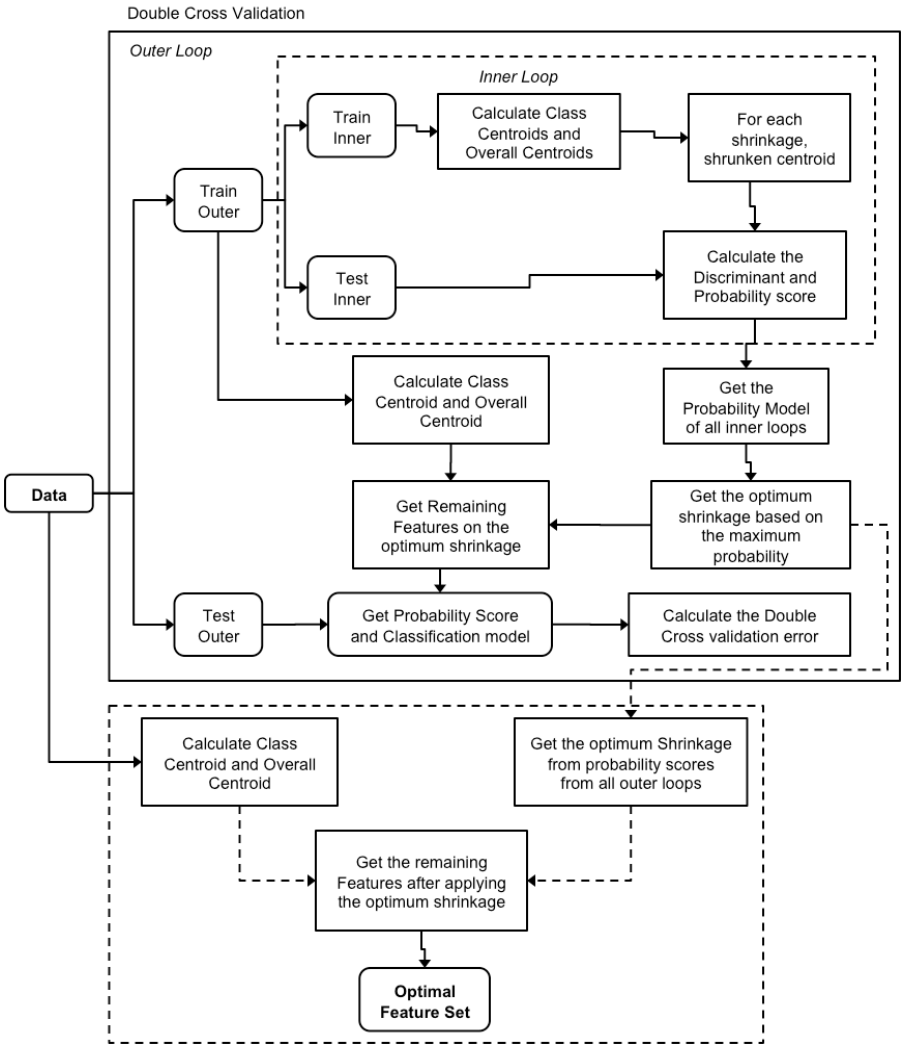

Figure 1. Double cross validation scheme for the Nearest Shrunken Centroids (NSC) algorithm. In the inner loops of the double cross validation scheme, the sum of the true class probability score at the respective shrinkage is calculated (maximum provides the optimal shrinkage). The final optimal feature set is selected using the shrinkage at the maximum of the median of the sum of the true class probability and the shrinkage plot after the double cross validation procedure. Performance of the classification is measured using optimal parameters in the outer loop by calculating the classification error rate on the outer loop training data set (double cross validation error).

2.2.3 Multivariate Support Vector Machine – Reduced Features Elimination (SVM-RFE)

A Support Vector Machine (SVM), originally proposed by Vapnik [23], is a multivariate supervised learning method that constructs a hyperplane which separates two groups in a given data set. Optimal separation between two classes is reflected by obtaining a hyperplane that has the largest distance to the nearest training data point of any class. A sample is viewed as *m*-dimensional vector, where *m* is the number of features. The goal of the SVM is to find a hyperplane with dimension *m*-1 that separates the vectors based on their respective classes. The hyperplane acts as a discriminant that assigns new data to a given class. SVM has been widely used and gained popularity for classification and prediction problems in medical research where the feature size far exceeds the available number of samples such as in microarray or mass spectrometry analyses [24-33]. Lately, approaches have been developed to adapt SVM for feature selection purposes [17, 34, 35]. In this study we use a linear SVM classifier combined with recursive feature elimination approach (RFE) for feature selection method as introduced by Guyon *et al.* [17]. This approach utilizes the weight vector **w**, which corresponds to the weight magnitude of features as the selection criterion during recursive feature elimination. The SVM-recursive feature elimination (SVM-RFE) procedure works as follows:

1. *Initially, using all the features in the training set, train the SVM classifier*
2. *Compute weight vector* $\boldsymbol{w}$
3. *Remove the feature with the lowest weight from the classification procedure*
4. *Train the SVM classifier using the remaining features*
5. *Repeat steps 2-4 until there is no remaining feature*

To obtain the optimal feature subset, we used a double cross validated support vector machine combined with recursive feature elimination (SVM-RFE). The optimal number of features is determined in the inner loop. Each time the feature with the lowest weight is eliminated, the classification error based on the new set of features is calculated. Each inner loop delivers a classification error for a given set of features and the rank of each feature is given by its weight. The optimal number of features is the smallest feature set that gives the minimum mean classification error. To select the optimum feature subset, a rank product procedure is applied to the feature rank lists produced in the inner loops. In the outer loop, the classification error of the optimal feature subset is computed using an independent test data set. The exact cross-validation scheme is shown in Figure S1 (Supporting Information).

2.2.4 Multivariate PCDA and PLSDA

PCDA and PLSDA take the relation between features into account in constructing new feature sets. Principal Component Analysis (PCA) constructs new features by finding linear transformations that best explain the variance in the data. PCA has been combined with Linear Discriminant Analysis (LDA) as the classifier applied on the PCA scores. This approach, originally proposed by Hoogerbrugge *et al* [19], has been used for feature selection and classification in biomedical research using mass spectrometry [36] or nuclear magnetic resonance data [37, 38].

PLSDA (Partial Least Squares - Discriminant Analysis) is a popular method in metabolomic studies [39-45] and was shown to be suitable for classification and discrimination in other applications [18, 46, 47]. It consists of a classical PLS regression analysis, where the response regressor is the class label. PLS components are built by trying to find a proper compromise between describing the data set and predicting the response. Further explanation and extensive assessment of this method can be found in Westerhuis *et al*. [48-50].

In our study we used a combination of double and single cross validation procedures for both PCDA and PLSDA. In double cross validation, the number of principal components (for PCDA) and the number of PLS components (for PLSDA) is optimized in the inner loop. At the end of each inner loop, a rank product procedure [20] is used to rank the features based on their discriminant coefficients obtained in the inner loop. The outer loop calculates the classification error from a model that uses the optimal number of PC/PLS components obtained in the inner loop and different numbers of features based on the ranked feature list. The feature size that gives the minimum classification error in the outer loops is selected as the optimal number of features. Based on this double cross validation procedure, the optimal number of principal components/PLS components and the optimal number of features are selected. To select the optimum feature sets, we utilize single cross validation separately from the double cross-validation procedure. In the cross validation loop, the model using the optimal number of components is built and the rank product of the discriminant coefficients of the features is calculated at the end using ranks obtained in each loops. The optimal feature set is selected from this ranked feature list whose size was given by the preceding double cross validation procedure. The complete scheme of PCDA and PLSDA in selecting the optimal feature set is shown in Figure S-2 (Supporting Information).

2.3 Evaluation criteria

We have performed the evaluation of the described approaches based on their performance as biomarker selection methods rather than as learning algorithms by measuring each algorithm's ability to construct an optimal feature set. In our case, where the discriminating features in the data sets are known, the optimal feature sets are supposed to contain only features related to the spiked peptides (true positives). True positives, false positives, true negatives and false negatives are subsequently identified in each feature set as proposed by a given method constructing a confusion matrix[51] (Table 2).

Truth \ by Methods	**Selected as optimal features (Positive)**	**Not selected (Negative)**
Spiked peptide-related features (Positive)	True Positive (tp)	False Negative (fn)
Non-spiked peptide-related features (Negative)	False Positive (fp)	True Negative (tn)

Table 2. The confusion matrix[51]. The columns correspond to features as predicted by a given method, while the rows correspond to the actual class of the features.

Several measures were calculated to compare and reveal the characteristics of the algorithms' performance (Table 3). Recall expresses the proportion of selected spiked-compound related features relative to all features that are related to the spiked peptides. Precision refers to the proportion of features that are related to the spiked peptides relative to all features selected by a given statistical method. The geometric mean accuracy (*g*-score) measures the ability of a method to classify both negative (not related to the spiked-peptides) features and positive (spiked-peptide related) features correctly. It assesses the overall performance of the feature selection methods, since it attributes the same importance to both true positive and true negative features. The *f*-score is a composite measure that concentrates on the correct classification of true positive features based on recall and precision. Recall calculates the proportion of the spiked-peptide related features that were selected as part of the optimal feature set relative to all spiked-peptide related features and assesses the effectiveness of an algorithm in identifying the true positive features. Precision is the proportion of the spiked-peptide related features amongst the selected feature sets and assesses the predictive power of a method. Recall and precision are balanced in the *f*-score when the *ß* constant parameter is set to 1 and is in favor of precision when $ß > 1$. In our work, we set *ß* equal to 1. We use the balanced *f*-score since we are interested in the correct identification of all spiked-peptide related features, which requires taking both recall and precision into account to the same extent.

Measure	Equation
Sensitivity = Recall = True Positive Rate (TPR)	$\frac{tp}{tp + fn}$
Precision	$\frac{tp}{tp + fp}$
Specificity = True Negative Rate (TNR)	$\frac{tn}{tn + fp}$
Geometric Mean Accuracy (*g*-score)	$\sqrt{TPR \cdot TNR}$
f-score	$\frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$

Table 3. Definition of the scores that were used to compare the performance of different feature selection methods.

Besides evaluating the optimal feature sets based on the aforementioned scores, performance of the methods can be measured based on several additional criteria: the classification error of the learning model that is built on its selected features, the complexity/number of the selected features and the stability of the selected feature subsets. Since the complexity of the learning model depends on the complexity of the feature set, the selected feature size has an impact on both performance and interpretability of the final model.

In biomarker discovery research, the size of the feature set determines the scale of subsequent experiments, such as the identification of selected peptides or proteins and their validation as biomarkers. Thus minimizing the number of false positives (maximizing precision) is more favorable than maximizing recall. In this paper we used designed data sets where the true positives are known. Therefore recall and precision can be calculated and compared across different algorithms. The confidence in selecting a set of biomarker candidates can be judged by the stability of the selected feature set upon repetition of the selection procedure. More confidence is achieved when the feature selection method gives similar feature sets across multiple repetitions of cross validation runs using different sample sets. Though the stability of the obtained feature sets cannot override the classification error with respect to new test samples, it is still a useful additional criterion for selecting an optimal feature subset from different models when the list of spiked-peptide related features is unknown.

3 RESULTS AND DISCUSSION

Six different statistical approaches were evaluated on LC-MS data sets from urine samples spiked with a range of peptides at different concentration levels as simulation of a biomarker discovery experiment. Figures 2 to 4 show the bar charts of the medians of the scores with the respective inter-quartile-ranges (IQRs) from 100 repetitions for each combination of statistical biomarker candidate selection method and data set. The median is used, since it gives robust measurements even when the distribution of the scores is not normal. Figure 2 shows the natural logs of the feature size (top) and the number of true positives (bottom) that are contained in the corresponding feature set. Figure 3 shows the recall (top) and precision (bottom) based on the number of true positive features found in the respective feature set. Figure 4 shows the *f*-scores and *g*-scores, which are a composite measure of recall, precision, and true negative rate. These scores were used to compare and assess the performance of each method with respect to sample size and class separation.

3.1 Comparison of individual methods

3.1.1 *t*-test and *mww*-test

Figures 2 and 3 show that neither the *t*-test nor the *mww*-test are capable of finding discriminating features between the two classes when sample size is low, independent of the magnitude of the concentration difference of the spiked peptides between the classes (data sets 1a and 2a, 6 samples per class). This improves markedly when increasing the number of samples per class to 15 (data sets 1c and 2c), where both tests are amongst the best performing methods. As expected from univariate methods, sample size has a strong influence on performance with a clear threshold between no or very poor performance for 6 samples per class and rather good performance for 15 samples per group. Figure 3 shows, however, that the high number of true positives in the results from data sets with 15 samples and a large class separation (see Figure 2, data set 1c) is accompanied by a relatively high number of false positives lowering precision in spite of a high recall. Despite a mediocre precision, univariate feature selection methods give the

highest *g*-score and *f*-score for this dataset (Figure 4, data set 1c), showing, that the trade-off between recall and precision results overall in the best performance. It is interesting to note that univariate tests give a lower recall but a higher precision when class separation decreases (data set 2c), which is primarily due to a 2-fold lower number of detected true positives (Figure 2, data set 2c) leading to a clearly decreased recall (Figure 3, data set 2c). The same tendency is observed for data sets 1b and 2b with 12 samples per class. The composite *f*-score and *g*-score show that the overall performance of univariate feature selection is mainly affected by sample size and slightly by class separation, and that standard univariate statistical methods perform as well as the more sophisticated multivariate or semi-multivariate methods with a class size of 15 samples in the case of spiked urine samples.

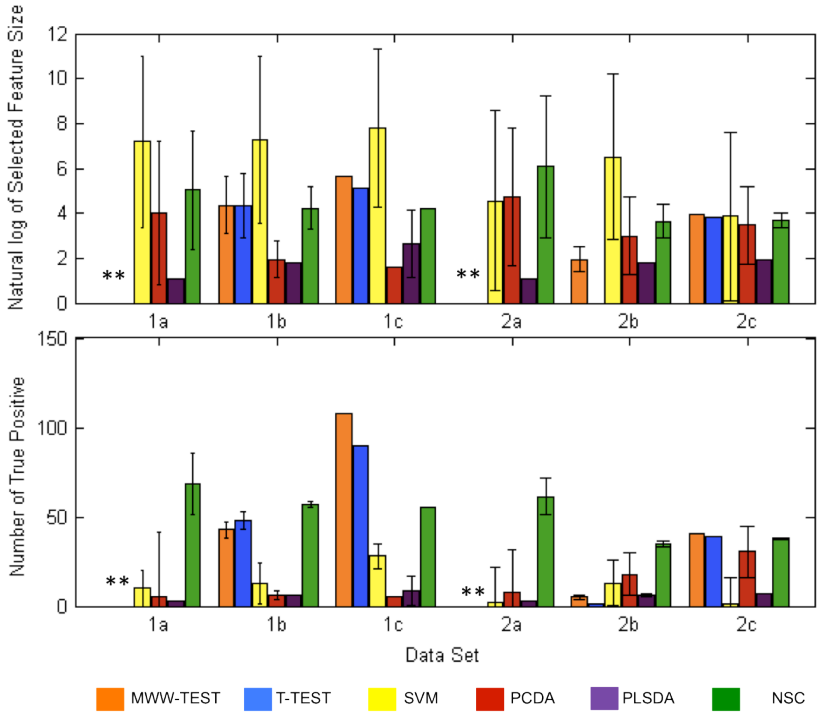

Figure 2. Bar charts of the median (± interquartile range) of the number of selected features (top) and the number of true positives (bottom) for each combination of feature selection statistical methods and data sets (see Table 1 for details concerning data sets). Results for the univariate tests (t-test and mww-test) on data sets 1a and 2a are denoted by **, since these methods selected no features at a sample size of 6. Univariate tests (t-test and mww-test) were performed once for data sets 1c and 2c comprising all available samples per class without repetition.

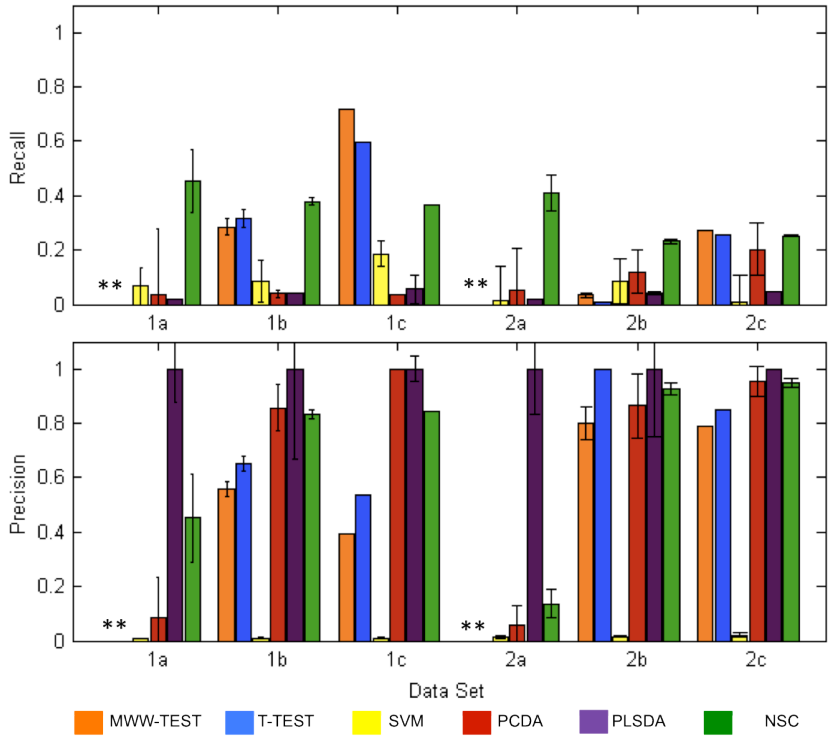

Figure 3. Bar charts of the median (± interquartile range) of recall (top) and precision (bottom) for each combination of feature selection statistical methods and data sets. Recall and precision are not available for the t-test and the mww-test on data sets containing 6 samples (denoted with **). Univariate tests (t-test and mww-test) were performed once for data sets 1c and 2c comprising all available samples per class without repetition.

3.1.2 NSC

NSC selected the highest number of true positives for the smallest sample size of 6 samples per class, however, at the expense of selecting also a fairly high number of features that are not related to the spiked peptides (Figure 2, data sets 1a and 2a). This leads to a high recall but intermediate precision when compared to the other algorithms (Figure 3, data sets 1a and 2a). Notably PLSDA outperforms NSC with respect to precision for data sets with a low number of samples due to a very low number of selected features that are not related to the spiked peptides (false positives). NSC remains amongst the better performing algorithms for data sets with a higher number of samples (data sets 1b or c and 2b or c) thus showing a fairly stable performance across all evaluated data sets, which is also reflected in rather similar *g*-scores and *f*-scores (Figure 4). NSC benefits from an increasing sample size when it comes to precision (Figure 3) since it selects less features that are not related to the spiked peptides, while the number of true positives decreases only slightly resulting in improved precision without sacrificing recall significantly. It is also noteworthy that the robustness of the statistical model improves with increasing sample size based on the reduced

interquartile range (see error bars in Figures 2 to 4). The difference of recall and precision between data sets with 12 and 15 samples was not significant (Figure 3, data set 1b vs 1c, and 2b vs 2c). Higher recall and slightly lower precision are observed for data sets with large class separation (1b and 1c) compared to those with small class separation (Figure 3, data sets 1b and 1c versus data sets 2b and 2c), which holds also for both the *f*-score and *g*-score (Figure 4).

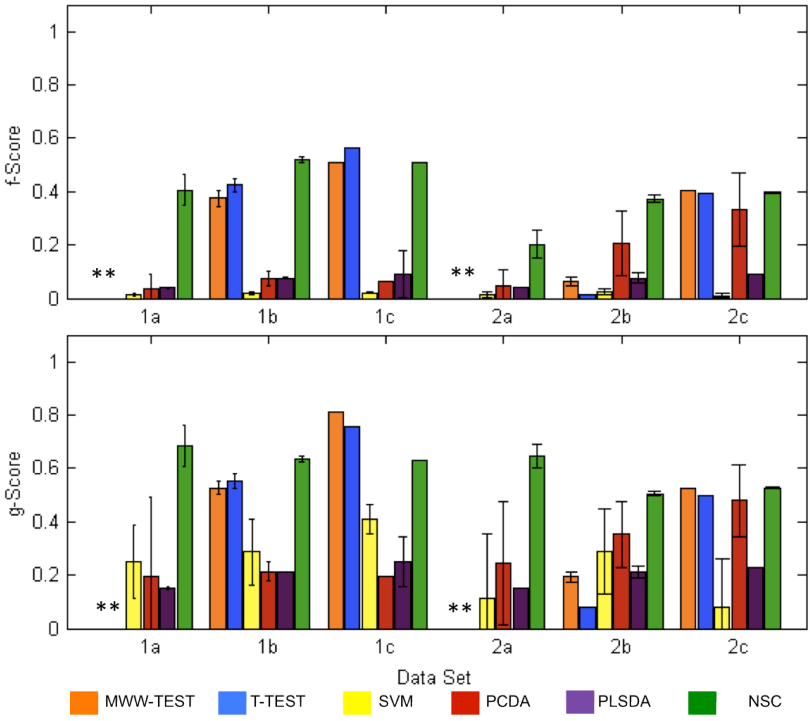

Figure 4. Bar charts of the median (± interquartile arnge) of f-score (top) and g-score (bottom) for each combination of feature selection statistical methods and data sets. The f-score and g-score are not available for the t-test and the mww-test on data sets containing 6 samples (denoted with **). Univariate tests (t-test and mww-test) were performed once for data sets 1c and 2c comprising all available samples per class without repetition.

3.1.3 SVM-RFE

SVM-RFE selected the highest number of features in almost all data sets, while the number of selected true positives was lower than for most of the other methods resulting in many false positives (Figure 2). Even though there is a trend to better performance as sample size increases and class separation is large (data set 1), scores remain low with a maximum recall of 0.2 and a maximum precision of 0.05 (Figure 3). The large number of selected features lowers precision and consequently the *f*-score (Figure 3 and 4). The *g*-score is also lower than for most other algorithms. It is the property of SVM that many correlated variables receive almost equal weights, which means that the weight of a

feature is not a very useful measure for feature selection. This could be the reason why the size of the selected feature set by SVM-RFE is rather large. Our results with peptide-spiked urine samples show that the SVM-RFE approach is not suitable for selection of biomarker candidates.

3.1.4 PCDA

For most of the studied data sets PCDA tends to select a low number of true positive features as well as a low number of features overall, however, often with considerable fluctuation resulting in low recall and high precision notably for large sample sizes (Figures 2 and 3, data sets 1b and c and 2b and c). PCDA may thus be considered a fairly 'conservative' approach to biomarker discovery, since the selected feature list has a relatively high content of true positive features. Contrary to NSC and the univariate tests (*t*-test and *mww*-test), PCDA tends to select fewer features from data sets with large class separation (data set 1), which is most pronounced for data set 1c. While the number of selected features is low (mean of 5.3 +/- 0.8), the selection contains essentially only true positive features (mean of 5.13 +/-0.8) resulting in a very high precision (mean of 0.97 +/- 0.06). This may be considered a positive characteristic of this approach when it comes to having to validate the selected biomarker candidates in large numbers of samples. It comes, however, at the expense that most of the true positives are missed using this statistical approach. Accurate selection of a low number of true positive features with low recall is reflected by the poor values of the composite measures *f*-score and *g*-score (Figure 4). Precision of the PCDA method is lower in data sets with low sample size irrespectively of class separation (Figure 3, data sets 1a and 2a), which makes the method adequate for biomarker candidate selection in data sets having a sample size equal or higher than 12 samples per group.

3.1.5 PLSDA

The most striking characteristic of PLSDA is the extremely high precision no matter the sample size or the class separation (Figure 3). This is advantageous in cases where subsequent biomarker validation is tedious and requires significant efforts. The stability of the model underlying feature selection increases with increasing sample size as shown by the reduced interquartile range. High precision comes at a price, however, since the number of selected true positives is small when compared to the number of expected true positive features related to the spiked peptides (Figure 2). The low number of selected true positives is likely due to the fact that PLSDA and PCDA exclude redundant discriminating features based on correlations between them. Since signals related to the spiked peptides are highly correlated, PCDA and PLSDA only select a few of them, which represent class separation well. Globally, PLSDA is not really affected by the strength of class separation in the data set, since the patterns of recall and precision in data sets 1 and 2 are comparable (Figure 3). When sample size increases, there is a pattern of slightly improving recall although recall remains overall low. Contrary to PCDA, PLSDA shows high precision also on data sets with a low sample size (Figure 3, data sets 1a and 2a) making this method more adequate for accurately selecting true

positive features in low sample size data sets such as 6 samples per sample group than PCDA.

3.2 Comparison between methods

While all methods benefit from a larger sample size, only some of them are affected by class separation. The univariate *t*-test, *mww*-test results are affected both by class separation and sample size (based on the comparison of *f*-score, *g*-score, recall and precision). They require furthermore a minimum sample size to function. Multivariate methods that use feature transformation prior to selecting a given feature set, such as PCDA and PLSDA, are not strongly affected by class separation or sample size. The performance of NSC is overall rather independent of class separation and sample size.

When the characteristics of the data set are profitable (large sample size and large class separation), univariate *t*-test and *mww*-test performed the best, since they assign most of the true positives within a reasonably sized total feature set. They are furthermore the fastest and simplest methods to use. Univariate methods fail, however, when sample size is small (e.g. 6 samples per class). Based on the *f*-score, *g*-score, recall and precision, the semi-multivariate NSC outperforms all other methods including multivariate methods in terms of feature selection, since it strikes the best balance between recall and precision, keeping both *f*-scores and *g*-scores high. The multivariate methods PLSDA and PCDA provide high-quality feature sets, which are reflected in high precision approaching 100% at the expense of a low recall. Globally speaking, NSC is applicable to all tested data sets and might be considered a good compromise when performing small-scale biomarker discovery studies.

Additional assessment criteria are the classification error rate or sum of true class probability in the case of NSC and the stability of the models. When repeating the calculation for a number of times, the result is trustworthier if all repetitions yield similar conclusions. To test this, we assessed the variability of feature selection performance across 100 repetitions based on *g*-score and *f*-score, variation of the classification error rate and the sum of true class probability in the case of NSC. Variability of *g*-score and *f*-score was measured using the IQR for a given data set. In general, there is a tendency of decreasing IQR for all approaches as the number of samples increases (see error bars in Figures 2 to 4), except for SVM-RFE, which may due to the poor performance of the approach resulting in quasi randomly selected feature set.

Variability of the classification error rate or the sum of true class probability in each cross validation loop reflects the stability of the model that is used as a classifier. Two different kinds of classification error can be derived from such a double cross validation scheme: the error in the inner loops, which determines the optimal values for the parameters and the classification error in the outer loops when using optimal parameter values. The inner loop classification error rate or sum of true class probability shows considerable dependency on the sample size of the data sets as shown in Figures S-3 to S-6 (Supporting Information). However class separation does not have an effect on inner loop classification error rates or on the sum of true class probabilities. The error in the outer loop averaged at 20% for data sets with large class separation (1a-c) and at 30% for data sets with small class separation (2a-c) independently of the applied method. The fluctuation of the inner loop classification error rate or the sum of true class probability

is due to different values for the optimal parameters determined in the inner loops obtained from varying training samples sets and can be assessed in a classification error or sum of true class probability against parameter plot. These plots indicate whether the minimum error, which determines the optimum parameter value, is located in a smooth/stable region. All plots that were used to determine the optimal parameters in this study are shown in s S-3 to S-6 (Supporting Information). The spread of the error decreases with increasing sample size showing that stability of the models increases with increasing sample number. The results show also that it is not justified to rely on a classification result from a model obtained from a training data set with only 6 samples per group, except in the case of the NSC algorithm.

To assess the variability of the selected feature sets delivered by each method, we compared the count of features that were selected at least once across 100 repetitions (unique features) relative to the count of features that were selected in each repetition (common features) (shown in Table S-1, Supporting Information). The stability of the feature set for the *mww*-test and the *t*-test for data sets 1c and 2c is not available because only one repetition was possible in these data sets due to the lack of a double cross-validation scheme for these methods. From this result, NSC produced the most stable feature set of all methods as shown by the high ratio of the number of common features to the number of unique features.

4 CONCLUSIONS

We have assessed different feature selection methods with respect to their capacity to deliver biomarker candidates from a number of well-controlled data sets that were obtained by LC-MS analysis of peptide-spiked human urine samples. Six widely used statistical methods were compared and their performance measured based on how well they find true positives (features that are related to the spiked peptides) and how well they avoid false positives (all other features) for data sets with different sample size and class separation. We derive five main conclusions from this study. (1) As expected, all methods benefit from a higher sample size. (2) Univariate methods and semi-multivariate methods are more sensitive to class separation, while multivariate methods (especially PLSDA) are hardly affected by class separation. (3) SVM-RFE performed poorly on all data sets with respect to selecting relevant features, showing that the weight vector is not a suitable criterion for feature selection/elimination. (4) True multivariate methods like PCDA and PLSDA aim at high precision by sacrificing recall, i.e. they are conservative with respect to selecting true positive features. PLSDA performs best for data sets with low sample size. (5) The semi-multivariate NSC strikes the best compromise between recall and precision regardless of sample size and class separation. Figure 5 provides an overview over the best performing biomarker candidate selection statistical methods based on the *f*-score for data sets with different class separation and sample size. This figure provides a summary concerning the choice of a given statistical method and should help practitioners to select the most suitable method for biomarker discovery studies.

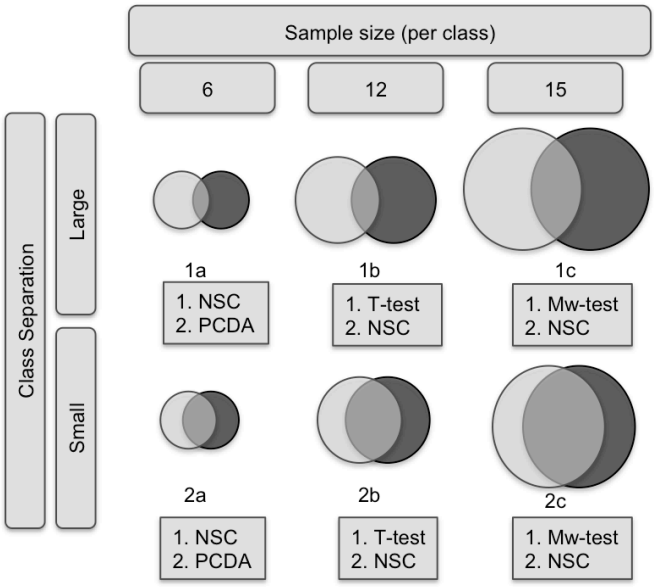

Figure 5. Overview over the two best performing feature selection statistical methods for data sets of different sample size and class separation based on the f-score. NSC shows its superior performance for data sets with 6 samples independent of class separation, while univariate tests rank on top when sample size increases to 15 samples per class.

Since biomarker discovery is usually intended to support clinical diagnosis, it is advantageous to obtain a discriminating feature set with a minimum number of false positives, and with the potential to classify new sets of samples correctly. Based on this criterion, PLSDA is a good choice due to its excellent precision. When additional discriminating features are required, for example to support pathway analysis, PLSDA may miss relevant features that could be informative. In this case, NSC provides a better compromise between recall and precision, with a higher number of true positives at a reasonable false positive rate. In cases where data from more samples are available (more than 15 samples per group/class in our case), univariate tests (*t*-test or *mww*-test with multiple testing correction) are able to identify biomarker candidates with high confidence. For classes with low sample numbers (six samples per class in our case) NSC has the highest potential to select biomarker candidates successfully. However our results show that there is a considerable danger when relying on results from data sets with such a small sample size, since classification models and the values for optimized parameters are prone to significant fluctuations making biomarker selection uncertain.

5 REFERENCES

1. Mischak, H.; Allmaier, G.; Apweiler, R.; Attwood, T.; Baumann, M.; Benigni, A.; Bennett, S. E.; Bischoff, R.; Bongcam-Rudloff, E.; Capasso, G.; Coon, J. J.; D'Haese, P.; Dominiczak, A. F.; Dakna, M.; Dihazi, H.; Ehrich, J. H.; Fernandez-Llama, P.; Fliser, D.; Frokiaer, J.; Garin, J.; Girolami, M.; Hancock, W. S.; Haubitz, M.; Hochstrasser, D.; Holman, R. R.; Ioannidis, J. P.; Jankowski, J.; Julian, B. A.; Klein, J. B.; Kolch, W.; Luider, T.; Massy, Z.; Mattes, W. B.; Molina, F.; Monsarrat, B.; Novak, J.; Peter, K.; Rossing, P.; Sanchez-Carbayo, M.; Schanstra, J. P.; Semmes, O. J.; Spasovski, G.; Theodorescu, D.; Thongboonkerd, V.; Vanholder, R.; Veenstra, T. D.; Weissinger, E.; Yamamoto, T.; Vlahou, A., Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med* **2010,** 2, (46), 46ps42.
2. Puntmann, V. O., How-to guide on biomarkers: biomarker definitions, validation and applications with examples from cardiovascular disease. *Postgrad Med J* **2009,** 85, (1008), 538-45.
3. Rifai, N.; Gillette, M. A.; Carr, S. A., Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **2006,** 24, (8), 971-83.
4. Saeys, Y.; Inza, I.; Larranaga, P., A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007,** 23, (19), 2507-17.
5. Smit, S.; Hoefsloot, H. C. J.; Smilde, A. K., Statistical data processing in clinical proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **2008,** 866, (1-2), 77-88.
6. Smit, S.; van Breemen, M. J.; Hoefsloot, H. C.; Smilde, A. K.; Aerts, J. M.; de Koster, C. G., Assessing the statistical validity of proteomics based biomarkers. *Anal Chim Acta* **2007,** 592, (2), 210-7.
7. Kohavi, R.; John, G. H., Wrappers for feature subset selection. *Artificial Intelligence* **1997,** 97, (1-2), 273-324.
8. Hilario, M.; Kalousis, A., Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform* **2008,** 9, (2), 102-118.
9. Baek, S.; Tsai, C. A.; Chen, J. J., Development of biomarker classifiers from high-dimensional data. *Brief Bioinform* **2009,** 10, (5), 537-46.
10. Datta, S.; Pihur, V., Feature selection and machine learning with mass spectrometry data. *Methods Mol Biol* **2010,** 593, 205-29.
11. Dougherty, E. R.; Hua, J.; Sima, C., Performance of feature selection methods. *Curr Genomics* **2009,** 10, (6), 365-74.
12. Van der Walt, C. a. B., E, Data characteristics that determine classifier performance. *17th Annual Symposium of the Pattern Recognition Association of South Africa* **2006**, 6.
13. Rubingh, C.; Bijlsma, S.; Derks, E.; Bobeldijk, I.; Verheij, E.; Kochhar, S.; Smilde, A., Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* **2006,** 2, (2), 53-61.
14. Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995,** 57, (1), 289-300.
15. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G., Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **2002,** 99, (10), 6567-72.
16. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G., Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. *Statistical Science.* **2003,** 203, (18), 104-117.
17. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V., Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002,** 46, (1), 389-422.
18. Barker, M.; Rayens, W., Partial least squares for discrimination. *Journal of Chemometrics* **2003,** 17, (3), 166-173.
19. Hoogerbrugge, R.; Willig, S. J.; Kistemaker, P. G., Discriminant analysis by double stage principal component analysis. *Analytical Chemistry* **1983,** 55, (11), 1710-1712.

20. Breitling, R.; Armengaud, P.; Amtmann, A.; Herzyk, P., Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **2004,** 573, (1-3), 83-92.
21. Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P., Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal Chem* **2008,** 80, (9), 3095-104.
22. Rosenling, T.; Slim, C. L.; Christin, C.; Coulier, L.; Shi, S.; Stoop, M. P.; Bosman, J.; Suits, F.; Horvatovich, P. L.; Stockhofe-Zurwieden, N.; Vreeken, R.; Hankemeier, T.; van Gool, A. J.; Luider, T. M.; Bischoff, R., The effect of preanalytical factors on stability of the proteome and selected metabolites in cerebrospinal fluid (CSF). *J Proteome Res* **2009,** 8, (12), 5511-22.
23. Vapnik, V., *Statistical Learning Theory*. Wiley-Interscience: 1998.
24. Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S., A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res* **2003,** 2, (2), 137-46.
25. Mao, Y.; Zhou, X. B.; Pi, D. Y.; Sun, Y. X., Constructing support vector machine ensembles for cancer classification based on proteomic profiling. *Genomics Proteomics Bioinformatics* **2005,** 3, (4), 238-41.
26. Jiang, Z.; Yamauchi, K.; Yoshioka, K.; Aoki, K.; Kuroyanagi, S.; Iwata, A.; Yang, J.; Wang, K., Support vector machine-based feature selection for classification of liver fibrosis grade in chronic hepatitis C. *J Med Syst* **2006,** 30, (5), 389-94.
27. Guo, J.; Deng, W.; Zhang, L.; Li, C.; Wu, P.; Mao, P., Prediction of prostate cancer using hair trace element concentration and support vector machine method. *Biol Trace Elem Res* **2007,** 116, (3), 257-72.
28. Mao, Y.; Zhao, X.; Wang, S.; Cheng, Y., Urinary nucleosides based potential biomarker selection by support vector machine for bladder cancer recognition. *Anal Chim Acta* **2007,** 598, (1), 34-40.
29. Lin, E.; Hwang, Y., A support vector machine approach to assess drug efficacy of interferon-alpha and ribavirin combination therapy. *Mol Diagn Ther* **2008,** 12, (4), 219-23.
30. Pham, T. V.; van de Wiel, M. A.; Jimenez, C. R., Support vector machine approach to separate control and breast cancer serum samples. *Stat Appl Genet Mol Biol* **2008,** 7, (2), Article11.
31. Webb-Robertson, B. J.; Cannon, W. R.; Oehmen, C. S.; Shah, A. R.; Gurumoorthi, V.; Lipton, M. S.; Waters, K. M., A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics* **2008,** 24, (13), 1503-9.
32. Henneges, C.; Bullinger, D.; Fux, R.; Friese, N.; Seeger, H.; Neubauer, H.; Laufer, S.; Gleiter, C. H.; Schwab, M.; Zell, A.; Kammerer, B., Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. *BMC Cancer* **2009,** 9, 104.
33. Zou, A. M.; Wu, F. X.; Ding, J. R.; Poirier, G. G., Quality assessment of tandem mass spectra using support vector machine (SVM). *BMC Bioinformatics* **2009,** 10 Suppl 1, S49.
34. Hermes, L.; Buhmann, J. M. In *Feature selection for support vector machines*, Proceedings 15th International Conference on Pattern Recognition, 2000. , 2000; 2000; pp 712-715.
35. Weston, J.; Mukherjee, S.; Chapelle, O.; Pontil, M.; Poggio, T.; Vapnik, V., Feature selection for SVMs. In MIT Press: 2000.
36. Hoefsloot, H. C.; Smit, S.; Smilde, A. K., A classification model for the Leiden proteomics competition. *Stat Appl Genet Mol Biol* **2008,** 7, (2), Article8.
37. Amato, U.; Larobina, M.; Antoniadis, A.; Alfano, B., Segmentation of magnetic resonance brain images through discriminant analysis. *Journal of Neuroscience Methods* **2003,** 131, (1-2), 65-74.
38. Lamers, R. J.; DeGroot, J.; Spies-Faber, E. J.; Jellema, R. H.; Kraus, V. B.; Verzijl, N.; TeKoppele, J. M.; Spijksma, G. K.; Vogels, J. T.; van der Greef, J.; van Nesselrooij, J. H., Identification of disease- and nutrient-related metabolic fingerprints in osteoarthritic Guinea pigs. *J Nutr* **2003,** 133, (6), 1776-80.

39. Ramadan, Z.; Jacobs, D.; Grigorov, M.; Kochhar, S., Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta* **2006,** 68, (5), 1683-1691.

40. Lv, Y.; Liu, X.; Yan, S.; Liang, X.; Yang, Y.; Dai, W.; Zhang, W., Metabolomic study of myocardial ischemia and intervention effects of Compound Danshen Tablets in rats using ultra-performance liquid chromatography/quadrupole time-of-flight mass spectrometry. *J Pharm Biomed Anal* **2010,** 52, (1), 129-35.

41. Liu, Y.; Huang, R.; Liu, L.; Peng, J.; Xiao, B.; Yang, J.; Miao, Z.; Huang, H., Metabonomics study of urine from Sprague-Dawley rats exposed to Huang-yao-zi using (1)H NMR spectroscopy. *J Pharm Biomed Anal* **2010,** 52, (1), 136-41.

42. Lan, K.; Zhang, Y.; Yang, J.; Xu, L., Simple quality assessment approach for herbal extracts using high performance liquid chromatography-UV based metabolomics platform. *J Chromatogr A* **2010,** 1217, (8), 1414-8.

43. Kim, H. K.; Saifullah; Khan, S.; Wilson, E. G.; Kricun, S. D.; Meissner, A.; Goraler, S.; Deelder, A. M.; Choi, Y. H.; Verpoorte, R., Metabolic classification of South American Ilex species by NMR-based metabolomics. *Phytochemistry* **2010,** 71, (7), 773-84.

44. Feng, B.; Wu, S. M.; Lv, S.; Liu, F.; Chen, H. S.; Gao, Y.; Dong, F. T.; Wei, L., A novel scoring system for prognostic prediction in d-galactosamine/lipopolysaccharide-induced fulminant hepatic failure BALB/c mice. *BMC Gastroenterol* **2009,** 9, 99.

45. Barba, I.; Garcia-Ramirez, M.; Hernandez, C.; Alonso, M. A.; Masmiquel, L.; Garcia-Dorado, D.; Simo, R., Metabolic fingerprints of proliferative diabetic retinopathy: an 1H-NMR-based metabonomic approach using vitreous humor. *Invest Ophthalmol Vis Sci* **2010,** 51, (9), 4416-21.

46. Boulesteix, A. L.; Strimmer, K., Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* **2007,** 8, (1), 32-44.

47. Chevallier, S.; Bertrand, D.; Kohler, A.; Courcoux, P., Application of PLS-DA in multivariate image analysis. *Journal of Chemometrics* **2006,** 20, (5), 221-229.

48. Westerhuis, J.; Hoefsloot, H.; Smit, S.; Vis, D.; Smilde, A.; van Velzen, E.; van Duijnhoven, J.; van Dorsten, F., Assessment of PLSDA cross validation. *Metabolomics* **2008,** 4, (1), 81-89.

49. Westerhuis, J.; van Velzen, E.; Hoefsloot, H.; Smilde, A., Discriminant Q2 (DQ2) for improved discrimination in PLSDA models. *Metabolomics* **2008,** 4, (4), 293-296.

50. Westerhuis, J. A.; van Velzen, E. J.; Hoefsloot, H. C.; Smilde, A. K., Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics* **2010,** 6, (1), 119-128.

51. Kohavi, R.; Provost, F., Glossary of terms: Special Issue on Application Machine Learning and Knowledge Discovery Process. *Machine Learning* **1998,** 30, 271-274.

Chapter 5

Summary and Future Perspectives

The development of new data analysis algorithms and the optimization of existing algorithms to support biomarker discovery research is a continuing challenge. It has to be continuously adapted to the variety of data produced by newly developed LC-MS techniques as one of the most important analytical tools for biomarker and biological knowledge discovery in the past decade. Improvements in data processing methods are crucial to strengthen a vast breakthrough in biomarker discovery research. This thesis presents recent supplements contributing to the continuous progress in achieving accurate methods and analysis for complex LC-MS data sets. A fundamental strategy in tackling time alignment problems raised from progressing complexity of research objectives and some guidelines in applying feature selection methods on low sample size data sets are discussed in this thesis.

As an introduction, **chapter 1** provides an overview of the main modules in label-free data processing pipelines including statistical analysis and validation. Here, the importance of the modules both as individual and as interconnected workflows is discussed while focusing on fundamental means of data processing and current challenges in biomarker discovery research using LC-MS of label free, shotgun proteomics data. In this chapter we showed that none of the modules in a data processing pipeline can be neglected. The choice and the combination of how these modules are connected must be adapted to the experimental and analysis design. Different types of LC-MS instruments generating varied data sets and different strategies in handling the corresponding data set are provided as additional information in this chapter.

Chapters 2 and **3** provide a strategy to overcome challenges in correcting retention time shifts across complex LC-MS chromatograms. Many time alignment algorithms, that were developed in their time were able to correct time shifts across different LC-MS runs only by using one-dimensional data such as total ion chromatogram (TIC) or base peak chromatogram (BPC), but failed to fix recent alignment problems when data sets and chromatograms became much more complex. Complex data sets contain more information as they contain more noise. Time alignment algorithms for data sets where thousands of biological compounds elute within a close time range require a step forward in handling and correcting the time shifts, such as: filtering mass traces with low noise but high signal, and taking this mass information into account in the time alignment function. In **chapter 2**, we proved that by considering the two mentioned important factors, the modified algorithm (COW-CODA) was able to align complex data sets, whereas the original algorithm Correlation Optimized Warping (COW) that used TIC failed to do so. We combined the original COW algorithm with the Component Detection Algorithm (CODA) to select high quality mass traces. These principals did not only work on COW, as shown in **chapter 3**, but also on other well-known and widely used time alignment algorithms: Parametric Time Warping (PTW) and Dynamic Time Warping (DTW). The result of the modified algorithms (DTW-CODA and PTW-CODA) performed strongly superior to their original. These modifications have improved the alignment quality of three different data sets: cancer serum, factorial design and urine. We showed that it is possible to add mass information in the benefit function of the three afore-mentioned algorithms, which work on different search spaces. For some other existing time alignment algorithms, which

originally developed to align one-dimensional data such as TIC/BPC, considering separated mass traces in the warping procedure, this may not be straightforward. However, in developing new time alignment algorithms, this principal must be incorporated.

Another aspect to consider when aligning a set of complex chromatograms is the choice of the reference chromatogram. In **chapter 2**, we proposed an approach to choose the best chromatograms based on the correlations between chromatograms in the data set. Even though the three algorithms investigated in this work are not significantly affected by the choice of reference chromatogram, it is certainly useful to always consider the reference selection prior to time alignment in various data sets using other time alignment algorithms. Reference selection is compulsory especially when the chromatograms in the data sets originate from high-variability biological samples, different analytical factors, different laboratories and/or instrumentation. Before the next data processing pipeline process can be carried out, the quality of the time alignment results must be checked. If present in the data set, internal standards should be considered as one of the validation criteria to examine the quality of individual algorithms. Since the internal standards are usually only a small representative of a complex chromatogram, or some data sets may not contain any internal standard at all, visualization of time alignment can be done simply by superimposing the reference chromatogram and sample chromatogram before and after alignment as this approach provides a better readout through the whole area of chromatogram. The validation of time alignment quality by visualization is based on a subjective judgment and requires more manual work. In complement, the overlapping peak area across all chromatograms gives a quantitative global approach to compare time alignment algorithms. Local/global visualization and the overlapping peak area provide a complete approach for validating and comparing the quality of time alignment algorithms.

While pre-processing methods deal with signal detection to recognize whether a signal belongs to a compound or noise, post-processing deals with giving this signal a further meaning: whether or not it is related to the disease stage. Discrimination processes have been developed in the statistical field and/or as part of classification problems. While other fields, which deal with classification or discriminant analysis, may have sufficient sample size, data sets in clinical research often suffer from an extreme high-dimensionality-small-sample (HDSS) problem. In **chapter 4** we present a comparative study of six different feature selection methods for LC-MS based proteomics and metabolomics biomarker discovery: *t*-test, Mann-Whitney-Wilcoxon-test (*mww*-test), Nearest Shrunken Centroid (NSC), linear Support Vector Machine – Recursive Features Elimination (SVM-RFE), Principal Component Discriminant Analysis (PCDA) and Partial Least Squares Discriminant Analysis (PLSDA) tested on data sets of urine samples that were spiked with a range of peptides at different concentration levels. In this chapter we investigated the behavior and performance of these methods using a data set with known discriminating compounds to facilitate the validation of the methods in selecting discriminating feature set. The statistical methods were applied to six data sets with different sample sizes and different class separations determined by the concentration level of spiked compounds. This study showed that all the methods benefit from high sample size data sets, fifteen samples per group in this study. Even

though these methods are widely used for feature selection problem in proteomics and clinical research, not all of them gave an acceptable result when tested on the spiked urine data sets. Univariate *t*-test and *mww*-test gave poor and unreliable results when applied on data sets with six samples per group, even though these univariate tests outperformed other methods when fifteen samples per group were used. Surprisingly, linear SVM-RFE gave inadequate performance in all six different data sets by selecting high numbers of features not related to the spiked compounds (leading to very low precision). PCDA performed rather unstable in data sets with six samples per group, but improved on data sets with larger sample size. Only PLSDA and NSC performed reasonably in the six data sets. PLSDA strikes excellent precision, while NSC provides a better compromise between recall and precision, with a higher number of true positives at a reasonable expense of false positives. In conclusion, it is worth noting that the results from feature selection methods require further validation when applied on data set with six samples per group. Therefore, one of the future works in this field should be dedicated to designing a strategy to validate the discriminating features in complement of cross-validation schemes for low sample size data sets with unknown discriminating compounds.

Even though the methods developed in this thesis are intended for proteomics research, they can be easily adapted to metabolomics research and data sets analyzed by different kinds of mass-spectrometry technique (GC-MS, LC-FT-MS, etc.). Furthermore, newly developed or enhanced algorithms are emerging within bioinformatics research groups. Therefore, future work should draw attention to the integration and the usability of these tools for a larger community. Rapid collaboration between groups and laboratory leads to high analytical variability within the data set. The data processing has to be able to distinguish this analytical variability within samples analyzed under different conditions from biological variability across different groups. Some advanced methods may require more computing and hardware power that may not be supported by the computers of users. In this case, the centralization and accessibility of computing power should be one of the important foci in developing bioinformatics platform for analyzing various biological experiments from diverse locations.

Appendix A

Samenvatting en Toekomstperspectief

De ontwikkeling van nieuwe dataanalysealgoritmen en de optimalisatie van bestaande algoritmen om onderzoek naar het vinden van biomarkers te ondersteunen is een voortdurende uitdaging. Het moet continu worden aangepast aan de variëteit van data die worden geproduceerd door nieuw ontwikkelde LC-MS-technieken als een van de belangrijkste analytische middelen voor biomarker- en biologische kennis in het afgelopen decennium. Verbeteringen in methoden voor dataverwerking zijn van cruciaal belang in een grote doorbraak in onderzoek naar nieuwe biomarkers. Dit proefschrift presenteert recente aanvullingen die bijdragen aan de continue vooruitgang in het realiseren van nauwkeurige methoden en analyse voor complexe LC-MS-datasets. In dit proefschrift worden een fundamentele strategie voor het oplossen van uitlijnenproblemen die ontstaan door de toenemende complexiteit van onderzoeksdoelstellingen en enkele richtlijnen voor het toepassen van variabelenselectiemethoden voor databestanden met weinig monsters beschreven.

Als inleiding geeft **hoofdstuk 1** een overzicht van de belangrijkste modules in label-free dataverwerkingslijnen waaronder statistische analyse en validatie. Hier wordt het belang van de modules als individuele en onderling verbonden werkstromen besproken waarbij de aandacht uitgaat naar fundamentele manieren van dataverwerking en huidige problemen binnen onderzoek naar biomarkerbepaling met behulp van LC-MS van label-free, willekeurige proteomicsdata. In dit hoofdstuk is aangetoond dat geen van de modules in een dataverwerkingslijn kan worden genegeerd. De keus en de combinatie van hoe deze modules zijn verbonden, moeten worden aangepast aan het experimentele en analyseontwerp. Verschillende soorten LC-MS-instrumenten die verschillende datasets en verschillende strategieën genereren bij het werken met het corresponderende data worden in dit hoofdstuk als aanvullende informatie gegeven.

Hoofdstukken 2 en **3** bieden een strategie voor het oplossen van problemen bij het corrigeren van verblijftijdverschuivingen over complexe LC-MS-chromatogrammen. Veel uitlijnalgoritmen die in hun eigen tijd zijn ontwikkeld, konden tijdverschuivingen corrigeren over verschillende LC-MS-uitvoeringen door het gebruik van eendimensionale data zoals Total Ion Chromatogram (TIC) of Base Peak Chromatogram (BPC), maar konden geen recente problemen oplossen als datasets en chromatogrammen veel complexer worden. Complexe datasets bevatten meer informatie naarmate ze meer ruis bevatten. Uitlijnalgoritmen voor datasets waar duizenden biologische samenstellingen binnen een korte tijd uitwassen, vereisen een stap voorwaarts in de aanpak en het corrigeren van tijdverschuivingen, zoals het filteren van massasporen met weinig ruis maar hoog signaal, en rekening houden met deze massa-informatie in de uitlijnen functie. In **hoofdstuk 2** hebben we aangetoond dat door rekening te houden met de twee genoemde, belangrijke factoren, het gewijzigde algoritme (COW-CODA) complexe datasets aankon, terwijl het oorspronkelijke algoritme Correlation Optimized Warping (COW) dat gebruikmaakte van TIC dit niet kon. We combineerden het oorspronkelijke COW-algoritme met het Component Detection Algorithm (CODA) om goede massasporen te selecteren. Dit werkte niet alleen bij COW, zoals aangetoond in **hoofdstuk 3**, maar ook bij andere bekenden en veelgebruikte uitlijnalgoritmen: Parametric Time Warping (PTW) en Dynamic Time Warping (DTW). Het resultaat van de gewijzigde algoritmen (DTW-CODA en PTW-CODA) presteerde veel beter dan hun origineel. Deze wijzigingen hebben de kwaliteit

van drie verschillende datasets verbeterd: serummonsters uit onderzoek naar baarmoederhalskanker, serummonsters uit factoriële onderzoeksopzetten en urinemonsters uit metabolomica-onderzoek. We hebben aangetoond dat het mogelijke is massa-informatie toe te voegen voor de verbeterde functie van de drie eerder genoemde algoritmen die op verschillende "search-space" werken. Voor bepaalde andere, bestaande uitlijnalgoritmen die oorspronkelijk zijn ontwikkeld voor eendimensionale data zoals TIC/BPC, rekening houdend met afzonderlijke massasporen binnen het warpingproces, kan dit gecompliceerd zijn. Maar bij de ontwikkeling van nieuwe uitlijnalgoritmen moet dit worden meegenomen.

Een ander belangrijk aspect bij het uitlijnen van een reeks complexe chromatogrammen is de keuze van het referentiechromatogram. In **hoofdstuk 2** hebben we een aanpak voorgesteld voor het kiezen van de beste chromatogrammen gebaseerd op de correlaties tussen chromatogrammen in het dataset. Hoewel de drie algoritmen die in zit werk zijn onderzocht niet significant worden beïnvloed door de keus van het referentiechromatogram, is het zeker handig de referentieselectie te overwegen voor uitlijnen in verschillende dataset met behulp van andere uitlijnalgoritmen. Referentieselectie is noodzakelijk, vooral als de chromatogrammen in de datasets van biologische monsters met hoge variabiliteit, verschillende analytische factoren, verschillende laboratoria en/of instrumentatie komen. Voordat het volgende proces voor dataverwerking kan worden uitgevoerd, moet de kwaliteit van de uitlijnentresultaten worden gecontroleerd. Indien aanwezig in het data moeten interne normen worden gezien als een van de validatiecriteria om de kwaliteit van individuele algoritme te onderzoeken. Omdat de interne normen vaak slechts een kleine representatie zijn van een complex chromatogram, of omdat bepaalde dataset mogelijk helemaal geen interne norm bevatten, is visualisering van uitlijnen mogelijk door het referentiechromatogram en monsterchromatogram voor en na alignment toe te voegen omdat dit een betere uitlezing biedt door het hele gebied van het chromatogram. De validatie van de kwaliteit van de uitlijnen via visualisatie is gebaseerd op een subjectieve beoordeling en vereist meer handmatig werk. Het overlappende piekoppervlak voor alle chromatogrammen geeft een kwantitatieve, algemene benadering om uitlijnalgoritmen te vergelijken. Lokale/algemene visualisatie en het overlappende piekoppervlak bieden een complete benadering voor het valideren en vergelijken van de kwaliteit van uitlijnalgoritmen.

Voorverwerkingsmethoden hebben te maken met signaaldetectie om te zien of een signaal tot een samenstelling of ruis behoort, en bij naverwerking krijgen deze signalen meer betekenis: of de signalen gerelateerd zijn aan het ziekteverloop. In het statistische veld en/of als onderdeel van een classificatieprobleem zijn er discriminatieprocessen ontwikkeld. Terwijl andere velden waarin men zich bezighoudt met classificatie of discriminantanalyse mogelijk voldoende monsters bevatten, kampen datasets in klinisch onderzoek vaak met extreme problemen in "high-dimensionality-small-sample" (HDSS; hoge dimensionaliteit en beperkte aantallen monsters). In **hoofdstuk 4** geven we een vergelijkende studie van zes verschillende variabelenselectiemethoden voor op LC-MS gebaseerde proteomica en metobolomica biomarkeronderzoek: *t*-test, Mann-Whitney-Wilcoxon-test (*mww*-test), Nearest Shrunken Centroid (NSC), lineair Support Vector Machine - Recursive Features Elimination (SVM-RFE), Principal Component Discriminant Analysis (PCDA) en Partial Least Squares

Discriminant Analysis (PLSDA) getest op datasets van urinemonsters met verschillende peptiden in verschillende concentraties. In dit hoofdstuk hebben we het gedrag en de prestatie van deze methoden onderzocht met behulp van een dataset met bekende discriminerende samenstellingen om de validatie van de methoden bij het selecteren van een discriminerende variabelen te vergemakkelijken. De statistische methoden werden toegepast op zes datasets met diverse aantallen monsters en verschillende klassescheidingen bepaald door het concentratieniveaus van spiked samenstellingen. Dit onderzoek toonde aan dat alle methoden profiteren van datasets met veel monsters, vijftien monsters per groep in dit onderzoek. Hoewel deze methoden algemeen worden gebruikt voor variabelenselectieproblemen in proteomica en klinisch onderzoek, gaven ze niet allemaal een acceptabel resultaat toen ze werden onderzocht op de datasets van spiked urine. Univariate *t*-test en *mww*-test gaven slechte en onbetrouwbare resultaten als ze werden toegepast op datasets met zes monsters per groep, hoewel deze univariate testen het beter deden dan andere methoden als er vijftien monsters per groep werden gebruikt. Verrassend genoeg gaf lineaire SVM-RFE ontoereikend prestatie op alle zes verschillende datasets door hoge aantallen kenmerken te selecteren die niet waren gerelateerd aan de spiked samenstellingen (wat leidde tot minimale precisie). PCDA was nogal instabiel in datasets met zes monsters per groep maar verbeterde zich bij meer monsters. Alleen PLSDA en NSC presteren redelijk binnen de zes datasets. PLSDA heeft uitstekende precisie, en NSC geeft een beter compromis tussen recall en precisie met een hoger aantal terecht positieven bij een redelijk aantal vals positieven. Concluderend kan worden opgemerkt dat het resultaat van variabelenselectiemethoden verdere validatie vergt als het wordt toegepast op datasets met zes monsters per groep. Daarom moet er in de toekomst in dit veld worden gewerkt aan het ontwerpen van een strategie om de discriminerende variabelen in aanvullende of kruisvalidatieschema voor datasets met weinig monsters en met onbekende discriminerende samenstellingen te valideren.

Hoewel de methoden die in dit proefschrift zijn ontwikkeld, zijn bedoeld voor proteomicaonderzoek, kunnen ze eenvoudig worden aangepast aan metabolomicaonderzoek en datasets die worden geanalyseerd door verschillende soorten massaspectometrietechnieken (GC-MS, LC-FT-MS, etc.). Bovendien komen er nieuwe ontwikkelde of verbeterde algoritmen uit onderzoeksgroepen van de bioinformatica. Het werk moet dus aandacht vestigen op de integratie en de bruikbaarheid van deze middelen voor grotere gemeenschappen. Snelle samenwerking tussen groepen en laboratoria leidt tot een grote analytische variabiliteit binnen het data. De dataverwerking moet deze analytische variabiliteit kunnen onderscheiden binnen monsters die onder verschillende omstandigheden van biologische variabiliteit worden geanalyseerd over verschillende groepen. Bepaalde geavanceerde methoden vereisen misschien meer rekenkracht en hardware die mogelijk niet wordt ondersteund door de computers van gebruikers. In dat geval moet de centralisatie en toegankelijkheid van rekenkracht een van de belangrijke focuspunten zijn in de ontwikkeling van een bioinformaticaplatform voor het analyseren van verschillende biologische experimenten vanuit verschillende lokaties.

Appendix B

SUPPORTING INFORMATION CHAPTER 2

1 THEORY OF COW-CODA

1.1 GLOSSARY

Mathematical notations:

Uppercase bold letter: matrix

Lowercase bold letter: vector

Italic upper and lowercase letter: constant element of matrices and vectors

Upper case superscripts: reference to the type (sample and reference) of chromatogram, except for retention time vector and LC-MS data.

Data transformation notation:

* denotes interpolated data

$\hat{s}$ denotes warped data

Variables:

N number of segment

m length of segment

t slack parameter

$\mathbf{F}$ cumulated correlation matrix with size of $(N+1)\times(L^R+1)$

$\mathbf{U}$ best position matrix with size of $(N+1)\times(L^R+1)$

$\mathbf{S}$ raw data of sample chromatogram in matrix of $(d \times L^S)$

$\mathbf{R}$ raw data of reference chromatogram in matrix of $(d \times L^R)$

$\mathbf{s}$ original retention time vector of the sample chromatogram with size of $(1 \times L^S)$

$\mathbf{r}$ original retention time vector of the reference chromatogram with size of $(1 \times L^R)$

$\hat{\mathbf{s}}$ warped retention time vector of the sample chromatogram with size of $(1 \times L^S)$

L^S length in number of elements of the sample retention time vector

L^R length in number of elements of the reference retention time vector

d number of mass chromatograms

a remainder after segmentation with length m of the sample chromatogram

b remainder after segmentation with length m of the reference chromatogram

$\mathbf{s}_i$ retention times of size $(1 \times m)$ from the vector $\mathbf{s}$ for a segment i

$\mathbf{r}_i$ retention times of size $(1 \times m)$ from the vector $\mathbf{r}$ for a segment i

$\mathbf{S}_i$ i^{th} segment of $\mathbf{S}$ based on $\mathbf{s}_i$ with size of $(d \times m)$

$\mathbf{S}_{ij}$ j^{th} mass trace and i^{th} segment of $\mathbf{S}$ based on $\mathbf{s}_i$ with size of $(1 \times m)$

$\mathbf{R}_i$ i^{th} segment of $\mathbf{R}$ based on $\mathbf{r}_i$ with size of $(d \times m)$

$\mathbf{R}_{ij}$ j^{th} mass trace and i^{th} segment of $\mathbf{R}$ based on $\mathbf{r}_n^{\text{idx}}$ with size of $(1 \times m)$

x_i^R	index of segment node *i* in the reference chromatogram,
x_i^S	initial index of segment node *i* in the sample chromatogram
$x_i^{S_v}$	allowed index of segment node *i* in the sample chromatogram
I_i^S	set of allowed positions for segment node *i* of the sample chromatogram
k_i	number of selected traces in segment *i*
$\mathbf{c}_i^S$	index vector where CODA is applied to sample chromatogram **S** for segment *i*
$\boldsymbol{J}_i$	set of indices that correspond to the indices of selected mass chromatograms for segment *i*
p_{ij}	*MCQ* product of segment *i* and mass trace *j*
$\mathbf{p}_i$	*MCQ* product vector of segment *i*
MCQ	implicit form to calculate *MCQ* (mass chromatographic quality) value using CODA as explained in Windig *et al.*[1]
f	implicit form of benefit function

1. Windig, W.; Phalp, J. M.; Payne, A. W., A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry* **1996,** 68, (20), 3602-3606.

1.2 Warping procedure

1. **Data characteristics.** We want to align a sample chromatogram **S** (size of $d \times L^S$) with retention time vector **s** (size of $1 \times L^S$) to a reference chromatogram **R** (size of $d \times L^R$) with retention time vector **r** (size of $1 \times L^R$). d is the number (size) of m/z values. The warping procedure warps **s** to match reference **r** based on the best matching **S** to **R,** resulting in $\hat{\mathbf{s}}$ having the same number of data points than **s** (size of $1 \times L^S$).

2. **Segmentation of the reference and sample chromatograms.** First, **s** and **r** are partitioned into segments of length m in the time dimension. We choose m in such a way that it divides L^S and L^R into the same number of segments N:

$$L^S = N \cdot m + a \quad (1)$$

$$L^R = N \cdot m + b \quad (2)$$

where a and b are $< m$.

We create a series of index vectors from $\mathbf{s}_i$ and $\mathbf{r}_i$ of size $(1 \times m)$ and, using the index vectors, a series of segmented matrices $\mathbf{S}_i$ and $\mathbf{R}_i$ with size of $(d \times m)$ for $0 \le i \le N-2$.

The last segment index vectors $\mathbf{s}_{N-1}$ and $\mathbf{r}_{N-1}$ have different lengths, since they take the remaining points after division of L^S and L^R by m, into account. The length of the last segment is obtained by addition of m to the differences of the segment length in order to assure a minimal length of m for the last segment:

$$\text{length}(\mathbf{s}_{N-1}) = m + L^S - N \cdot m = m + a \quad (3)$$

$$\text{length}(\mathbf{r}_{N-1}) = m + L^R - N \cdot m = m + b \quad (4)$$

The last segment matrix $\mathbf{S}_{N-1}$ and $\mathbf{R}_{N-1}$ is obtained similarly with dimensions of $(d \times (m+a))$ and $(d \times (m+b))$, respectively. Each segment border refers to a node. Since the last index of a segment equals the first index of the next segment, there are N+1 segment borders, or rather N+1 nodes.

3. **Segment node indices and allowed changes in search space for optimal time alignment.** The index of segment nodes x_i^R for chromatogram **R,** where $i = 0, \cdots, N$ of segment $\mathbf{r}_i$ in chromatogram **R,** are calculated as:

$$x_i^R = i \cdot m + 1; \text{ for } i = 0, \ldots, N-1 \text{ and } x_N^R = L^R \quad (5)$$

Initially the segment nodes x_i^S of the chromatogram **S** are segmented similarly:

$$x_i^S = i \cdot m + 1; \text{ for } i = 0, \ldots, N-1 \text{ and } x_N^S = L^S \quad (6)$$

Indices of segment nodes x_i^S of chromatogram **S** are allowed to vary ($x_i^{S_V}$) during the warping procedure in a way that the changed segment length of $\tilde{s}_i$ must be within *m*±*t* as defined by the new node indices $x_i^{S_V}$ and $x_{i+1}^{S_V}$, where *t* is the slack parameter which is given in number of points. This variation defines also the set of node indices, where segment shrinking and stretching are allowed. Besides this there are three additional constraints, which limit the search space to find the optimally warped segments: $x_0^{S_V} = 1$, $x_N^{S_V} = L^S$ and

$$x_0^{S_V} = 1 < x_1^{S_V} < \cdots < x_{N-1}^{S_V} < x_N^{S_V} = L^{S_V} \tag{7}$$

The set I_i^S of possible nodes is for segment *i* defined in the following way. First define the set of lower boundaries by

$$I_{i_{\text{lower}}}^S = \left[i \cdot (m-t); i \cdot (m+t)\right]; \quad i = 1, \cdots, N-1, \tag{8}$$

and then the set of upper boundaries by

$$I_{i_{upper}}^S = \left[L^S - \left(N-i\right) \cdot \left(m+t\right); L^S - \left(N-i\right) \cdot \left(m-t\right)\right]; \quad i = 1, \cdots, N-1. \tag{9}$$

Note that the intervals will only contain integer values since the nodes can attain only integer values. The allowed nodes $x_i^{S_V}$ are now provided by the intersection of the intervals in (9) and (10):

$$x_i^{S_V} \in I_i^S = I_{i_{\text{lower}}}^S \cap I_{i_{\text{upper}}}^S; \quad i = 1, \cdots, N-1; \text{ where } I_0^S = \{0\} \text{ and } I_N^S = \{L_S\} \tag{10}$$

Warping segment $\mathbf{s}_i$ onto $\tilde{s}_i$ such that it matches with segment $\mathbf{r}_i$ is finding the optimal nodes $x_0^{S_V}$, $x_1^{S_V}$,..., $x_N^{S_V}$ in formula (10) that maximize the cumulative benefit function $f(\mathbf{S}_i^*, \mathbf{R}_i)$, in such away that the length of the interval $[x_i^{S_V}; x_{i+1}^{S_V}]$ is within ($m \pm t$). Here $\mathbf{S}_i^*$ is obtained from and $\mathbf{S}_i$ with linear interpolation to $\mathbf{R}_i$, to make as many intensity points as $\mathbf{R}_i$.

$f(\mathbf{S}_i^*, \mathbf{R}_i)$ is the value of the benefit function calculated from $\mathbf{S}_i^*$ and $\mathbf{R}_i$, which is used as criterion to find the optimum pathway between the reference and sample chromatograms, as explain later in the coming section (formula 14).

Segment-wise trace selection using CODA.

In the COW-CODA algorithm, for each segment *i*, k_i mass chromatograms are selected using CODA, with $0 \le k_i \le 30$ and $i = 0,\cdots N-1$. First *MCQ* values for all segments with nodes defined by equation (5) are calculated for the reference chromatogram. Since the segment length can vary in the sample chromatogram and the calculation of *MCQ* values for each pair of allowed segment index positions between each $\boldsymbol{I}_i^{\mathrm{S}}$ and $\boldsymbol{I}_{i+1}^{\mathrm{S}}$ will result in increased calculation time, the *MCQ* values for sample chromatograms are calculated in larger intervals than the original segment nodes as defined by equation (6) as follows:

$$\mathbf{c}_i^{\mathrm{S}} = \left[\min(I_i^{\mathrm{S}}), x_{i+1}^{\mathrm{S}} + t\right] \quad \text{for } i = 1,\cdots,N-2 \tag{11}$$

and for the last segment:

$$\mathbf{c}_{N-1}^{\mathrm{S}} = \left[\min(I_{N-1}^{\mathrm{S}}), (L^{\mathrm{S}})\right] \tag{12}$$

The presented boundary $\mathbf{c}_i^{\mathrm{S}}$ does not cover the entire interval of segment indices, but was chosen after trying several combinations of intervals.

After selection of segment nodes, *MCQ* values are calculated for both chromatograms and the product of *MCQ* values for each mass trace in every segment is determined. For a given segment:

$$p_{ij} = MCQ(\mathbf{S}_{ij}[\mathrm{c}_i^{\mathrm{S}}]) \cdot MCQ(\mathbf{R}_{ij}) \quad \text{for } j = 1,\cdots,d \text{ and } i = 0,\cdots,N \tag{13}$$

For segment *i* mass traces corresponding to the first *n* highest values in $\mathbf{p}_i$ with a lower limit of 0.77[2] are selected $\boldsymbol{J}_i$ and further used to determine the value of the benefit function. Figure S-2 shows an example of a mass trace (m/z = 466) with an MCQ product of 0.59 demonstrating that mass traces at the lower *MCQ* limit are still of high quality.

4. Now define the cumulative benefit function f in the nodes $x_0^{S_v}$, $x_1^{S_v}$, …, $x_N^{S_v}$ by

$$f\left[x_0^{S_v}, x_1^{S_v}, \ldots, x_N^{S_v}\right] = \sum_{i=0}^{N-1} f_i\left[x_i^{S_v}; x_{i+1}^{S_v}\right], \tag{14}$$

with the benefit function $f_i\left[x_i^{S_v}; x_{i+1}^{S_v}\right]$ for segment i =0,1,2,…., $N-1$ given by

$$f_i\left[x_i^{S_v}; x_{i+1}^{S_v}\right] = \prod_{j \in J}^{j} \rho\left(\mathbf{S}_{ij}^{*}\left[x_i^{S_v}; x_{i+1}^{S_v}\right], \mathbf{R}_{ij}\right).$$

The nodes $x_0^{S_v}$, $x_1^{S_v}$, ..., $x_N^{S_v}$ in (10) that maximizes the cumulative benefit function in (14) under the restriction that $m - t \le x_{i+1}^{S_v} - x_i^{S_v} \le m + t$ holds for all i =0,1,2,…., $N-1$ are the selected nodes for the time alignment. Thus the selected nodes are defined by

$$\hat{\mathbf{x}} = \operatorname{argmax}\left\{\begin{array}{l} f\left[x_0^{S_v}, x_1^{S_v}, \ldots\ldots, x_N^{S_v}\right]; \\ x_i^{S_v} \in I_i^S \wedge m - t \le x_{i+1}^{S_v} - x_i^{S_v} \le m + t, \\ i = 0,1,\ldots, N-1 \end{array}\right\} \qquad (15)$$

Retention times of these selected nodes are adapted to match the retention time of the corresponding nodes x_i^R . The warped retention time vector $\hat{\mathbf{s}}$ is determined by segment wise linear interpolation to x_i^R in order to keep the original sampling rate of **s.**

FIGURES SUPPORTING INFORMATION

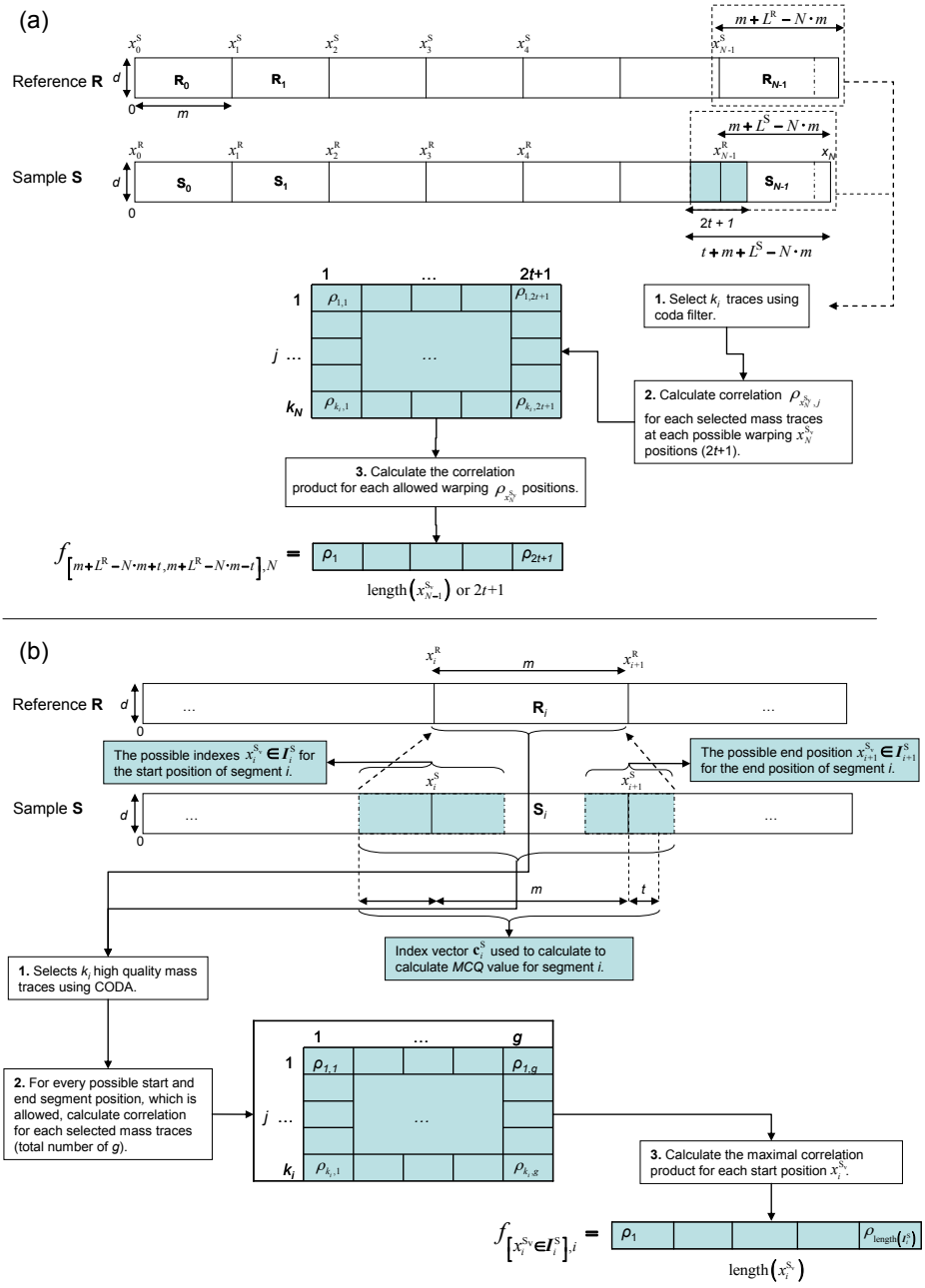

Figure S-1. (a) Alignment of the last segment using CODA selected mass chromatograms. **R** is the reference chromatogram and **S** is the sample chromatogram to be aligned. **S** and **R** (containing d mass chromatograms) are divided into N segments of length m. t is the slack parameter, except for the last segment ($\mathbf{S}_N$ and $\mathbf{R}_N$)**,** because this segment includes the remaining data points. The CODA

algorithm is applied to segments $\mathbf{S}_N$ and $\mathbf{R}_N$, resulting in k_N (where $k_N \leq 30$) selected mass chromatograms having $MCQ \geq 0.77$. There are $(2 \cdot t + 1)$ possible warping positions of segment S_N, which can be chosen to match the starting position of R_N. The correlation $\rho_{x_N^{S_y}}$ is obtained from $\prod_{j=1}^{k_N} \rho_{x_N^{S_y},j}$ between allowed segment positions of sample and the corresponding segment of the reference chromatogram of the k_N CODA selected mass traces which will be stored in matrix **F**.

(b) Alignment procedure for the remaining segments. The CODA algorithm was applied to $\mathbf{S}_i$ and $\mathbf{R}_i$ for $0 \leq i \leq N\text{-}1$, which are the i^{th} segments, and aligned following the $(i+1)^{th}$ segments. $\boldsymbol{I}_i^S$ set contains the possible starting node positions of segment i. $\boldsymbol{I}_{i+1}^S$ is the possible position of node $(i+1)$, whose cumulated correlation product from the last segment to segment i were calculated. **(**Figure S-2; b-2**)**. There will be g allowed positions with length $m \pm t$ between nodes $(i+1)$ and i. **(**Figure S-2; b-3**).** The correlation product $\prod_{j=1}^{k_i} \rho_{x_i^{S_V},j}$ between every allowed node positions $x_i^{S_V}$ and $x_{i+1}^{S_V}$ is added to the cumulated correlation product of $x_{i+1}^{S_V}$ (node $i+1$). For each position of node i the position of node $i+1$, which gives the maximum cumulated value, is chosen as the best warping position for this node i and is saved in matrix **U** at the location of $\left[x_i^{S_V}, i\right]$, while its cumulated value is saved in matrix **F** at the same location. The procedure starts with the last segment (see Figure 2a) and continues in backward direction, as described in Figure b, until reaching the first segment. The best warping position is then obtained from **U** starting from $u_{1,1}$.

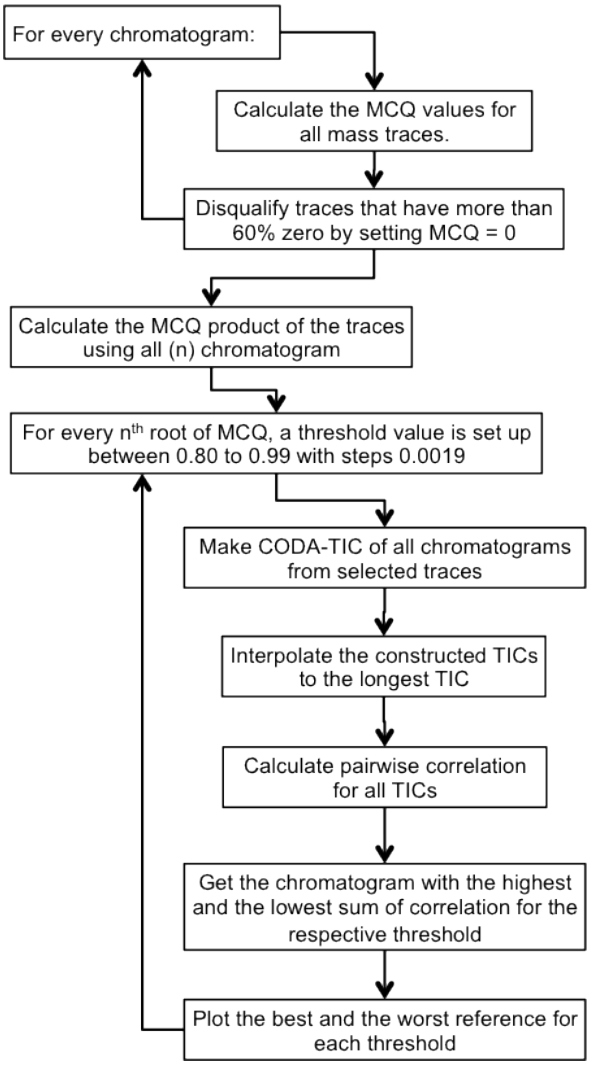

Figure S-2. Flow-chart to chose the best and worst reference chromatogram in a given LC-MS dataset. The algorithm is based on correlation between the TIC reconstructed from CODA-selected mass chromatograms having the n^{th} square root of the *MCQ* product of all (*n*) selected mass chromatograms with values above a given threshold (between 0.80 and 0.99).

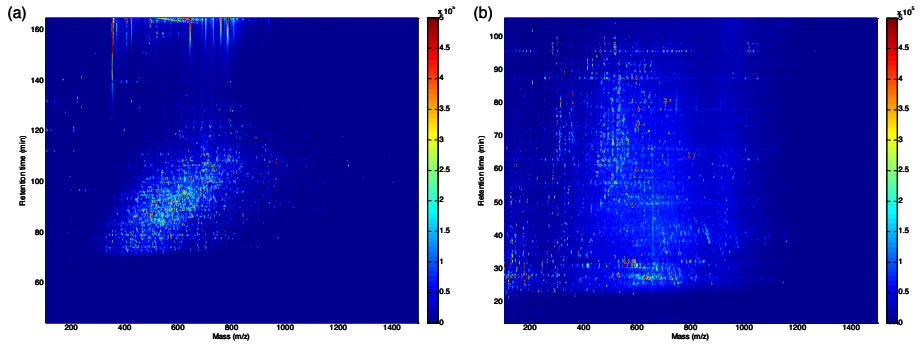

Figure S-3. Image plot with intensity coloration of depleted and trypsin-digested serum sample (a) and acid-precipitated urine (b).

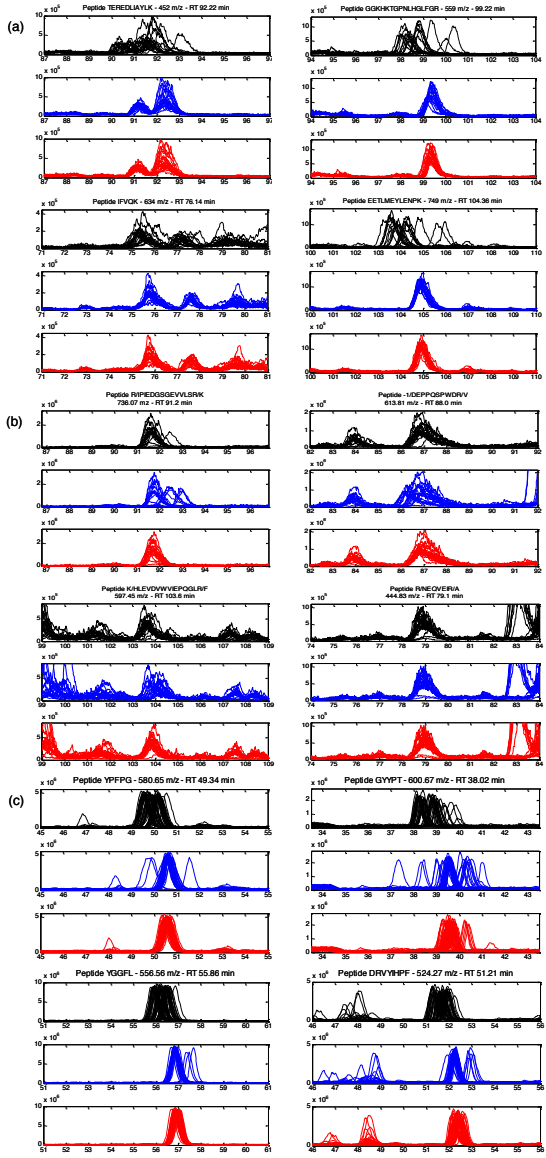

Figure S-4. Extracted ion chromatogram of internal standard peptides. (a) Dataset 1 (cervical cancer serum) comprised of 20 chromatograms, (b) Dataset 2 (factorial design serum) comprised of 19 chromatograms, and (c) Dataset 3 (acid-precipitated urine) comprised of 50 chromatograms. Each peptide is presented before alignment (top/black), after alignment by COW-TIC (middle/blue), and after alignment by COW-CODA (bottom/red). These results were obtained using the worst references, which were chromatograms with the numbers 19, 19 and 45 for Dataset 1, 2 and 3, respectively. Significant misalignments are observed for Dataset 2 and 3 when using the COW-TIC algorithm, while COW-CODA algorithm resulted in well aligned peak clusters.

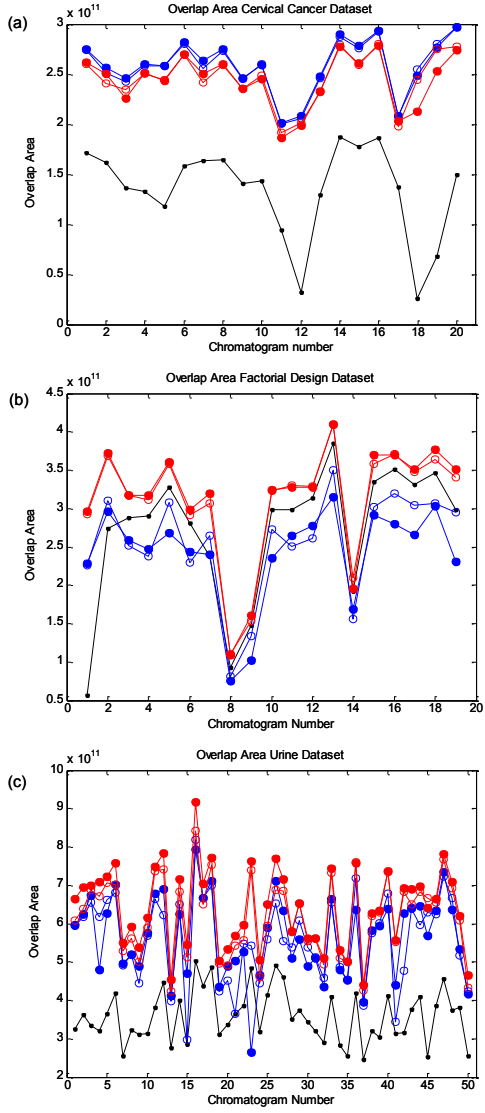

Figure S-5. Calculated peak area overlap for Dataset 1 (a), Dataset 2 (b), and Dataset 3 (c) with respect to the chosen reference chromatogram (best [red-filled circle] or worst [red-empty circle]). Calculated peak areas are compared within each dataset before warping (black), after warping using COW-TIC (bleu) and COW-CODA (red). The overlapping peak area does not change considerably with respect to the chosen reference chromatogram, showing that the COW-CODA algorithm finds the optimal alignment independent of the reference chromatogram.

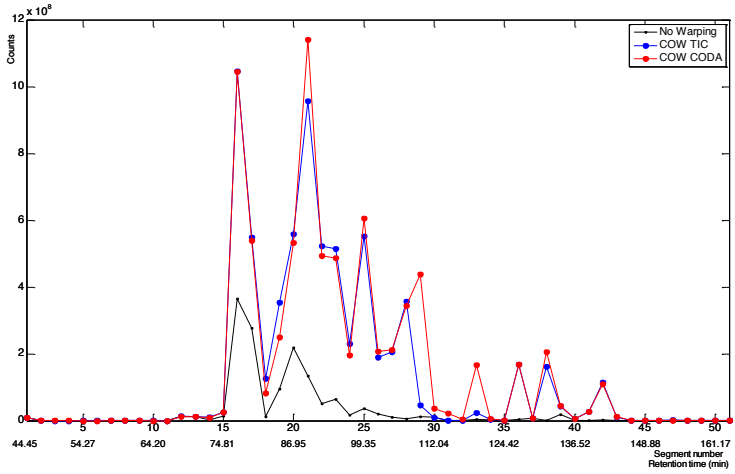

Figure S-6. Calculated peak area overlap for chromatograms 1 and 14 (best reference) for Dataset 2 (factorial design) for each segment. Calculated peak areas are compared for each segment before warping (black), after warping using COW-TIC (blue) and COW-CODA (red). While most segments show that the two time alignment methods perform equally well, it is obvious that COW-CODA is substantially superior to COW-TIC for 21, 25-27, 29-31, 33, 38.

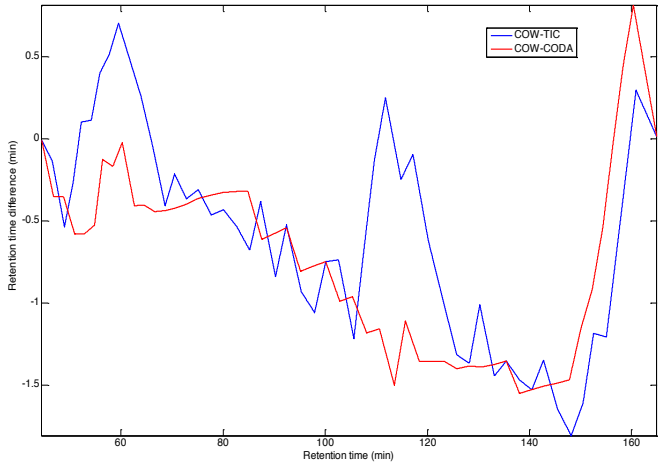

Figure S-7. Warping function obtained with COW-TIC (blue) and COW-CODA (red) on chromatograms 1 and 14 from Dataset 2 (factorial design). The region between 110 and 125 min shows large differences between the retention time shifts made by COW-TIC and COW-CODA. The better performance of the COW-CODA algorithm in this region is also observed in Figure S-6.

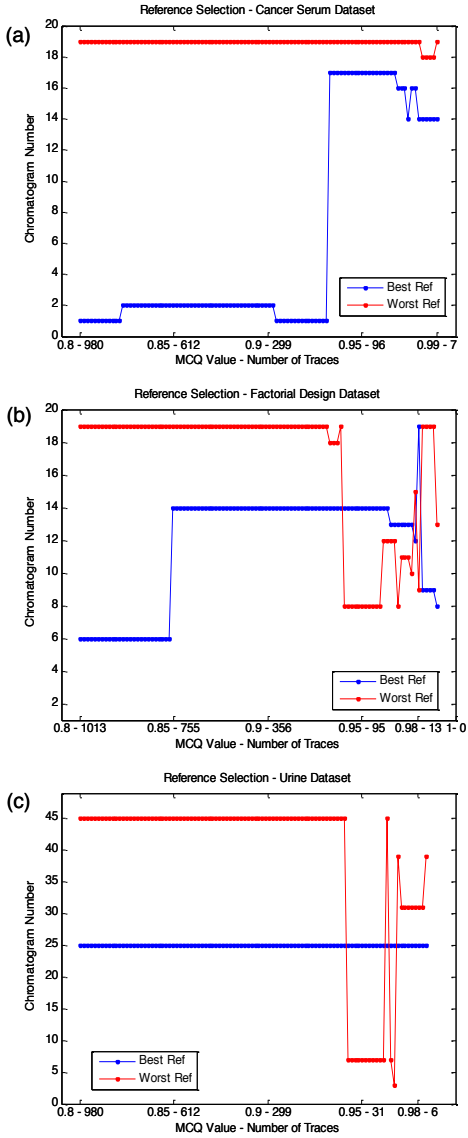

Figure S-8. Plots of *MCQ* values against chromatograms with the highest correlation (blue) and the lowest correlation (red) for Dataset 1 (a), Dataset 2 (b) and Dataset 3 (c). The x-axis represents the *MCQ* value and the number of selected traces to construct the CODA-TIC. A chromatogram is selected as the best or the worst reference if it has the highest or lowest correlation in the majority of the cases. Chromatogram 2 was selected as the best and chromatogram 19 as the worst reference for Dataset 1 (a). Chromatogram 14 was selected as the best and chromatogram 19 as the worst reference for Dataset 2 (b). Chromatogram 25 was selected as the best and chromatogram 45 as the worst reference for Dataset 3 (c).

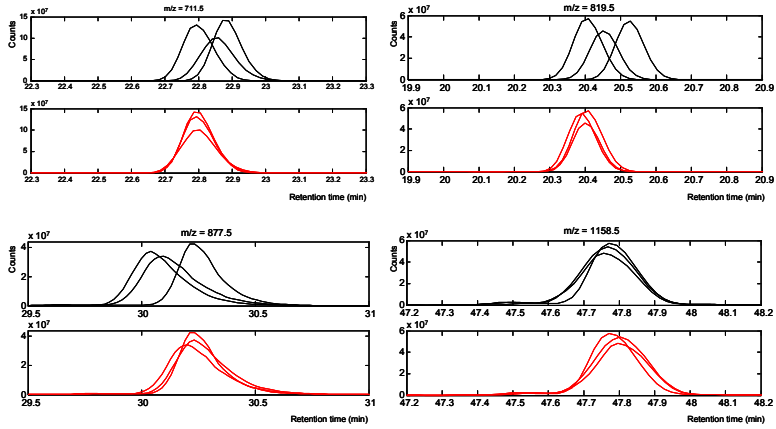

Figure S-9. Selected ion chromatograms of peaks of 3 chromatograms obtained from the TRANCHE database, "Pepper, a platform for experimental Proteomic Patter Recognition: Peak lists in mzxml format" project and acquired using a ThermoFinnigan LTQ FT instrument. The upper black traces show the original retention time, the lower traces (red) were obtained after applying COW-CODA.

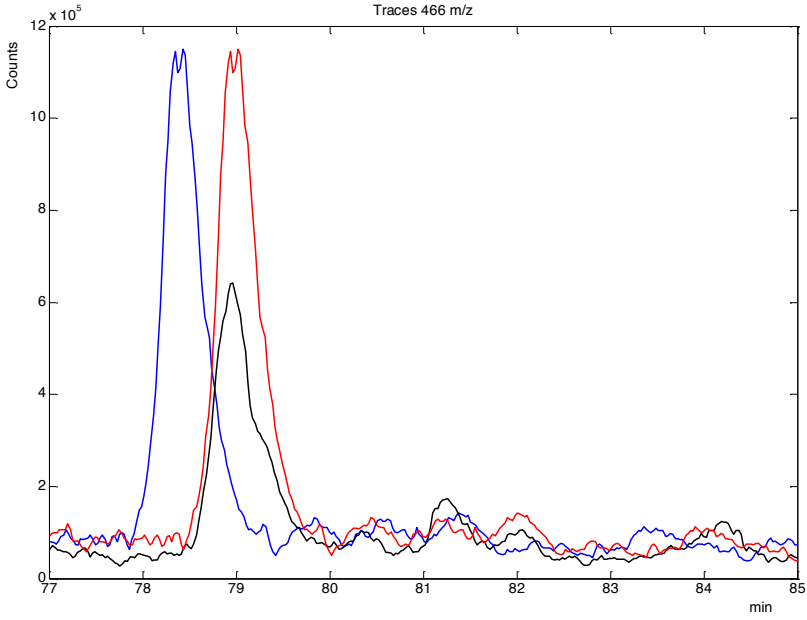

Figure S-10. Example of selected mass traces at the lowest threshold (*MCQ* product of 0.59) for the reference (chromatogram 25, black) and a sample chromatogram (chromatogram 1; blue: before alignment; red: after alignment with COW-CODA) from the urine dataset for segment 32.

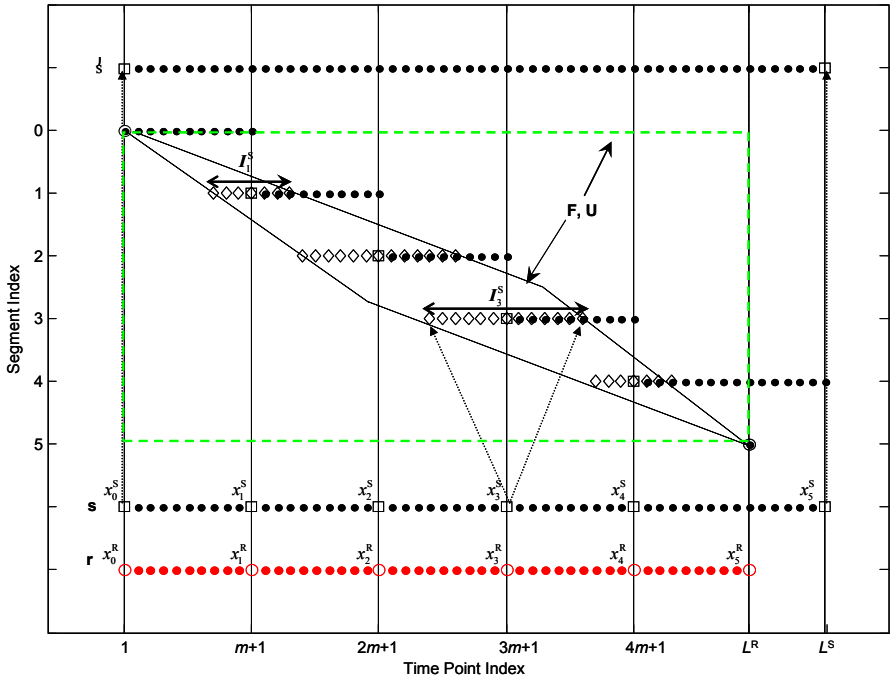

Figure S-11. Example of the warping procedure showing the diamond structure of the F and U matrices (green box with broken line) for m = 10, t = 3, and N = 5. Interval $\boldsymbol{I}_i^S$ = [8;14] shows the possible positions of node x1, and interval $\boldsymbol{I}_3^S$ = [25;37] shows possible positions for node x_3^S.

Appendix C

Supporting Information Chapter 3

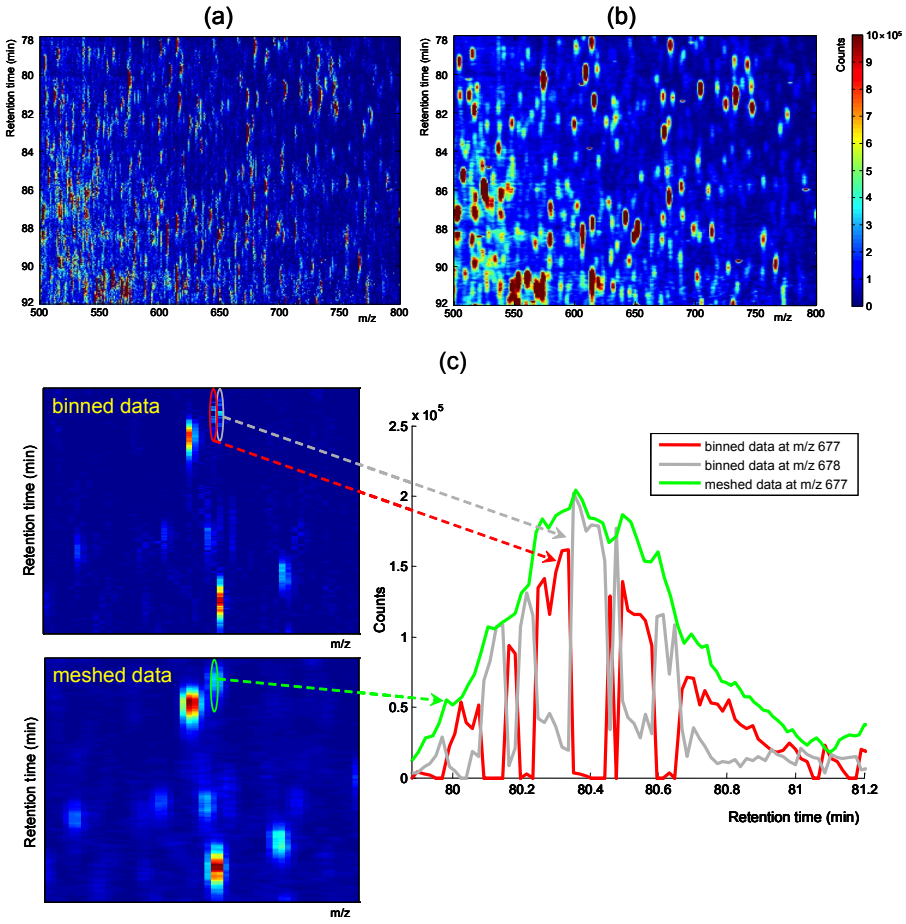

Figure S-1. Raw centroided ion trap single stage LC-MS image of depleted, trypsin-digested human serum obtained after binning (summing up intensity across 1 amu intervals having borders of 0.5 m/z for each integer m/z value) (a) and after applying a two-dimensional Gaussian filter using the same data reduction in the m/z dimension as for binning (b). Binning results in noisy data, which lead to a higher accumulated error in the score of the benefit function of the time alignment algorithms, in contrast to smoother data obtained with Gaussian smoothing. The contribution of this noise to the benefit function is shown in panel (c) representing part of an LC-MS image highlighting a peak using the extracted ion chromatogram, where the highest intensity of the peak in centroid data fluctuates between the border of the bins. Binned data will result in random fluctuations, which will be different for each chromatogram, and will lead to a lower correlation and a higher error in the benefit function of the time alignment algorithms.

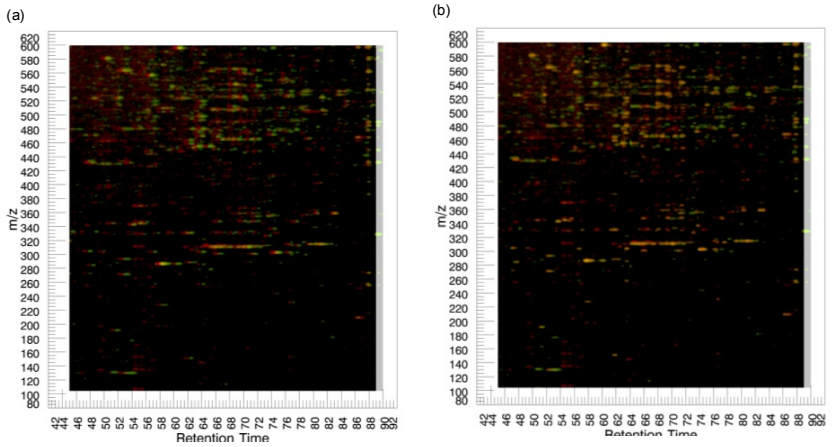

Figure S-2. Two-dimensional image view of the overlaid raw data in a section of the label-free single stage LC-MS analysis of two urine samples (5082628 and 5082602) before alignment (a) and after applying DTW-CODA (b) for m/z between 80-620 amu and the retention time between 42-92 min. The intensity of the reference chromatogram image is colored between red and black from high to low values, the sample chromatogram image is colored between green and black from high to low values. The overlay of the two chromatogram is transparent, which results in a yellow color if a pixel has the same intensity in the two chromatograms providing an image similar to two-dimensional fluorescence difference gel electrophoresis. Common peaks are superimposed after warping (indicated in yellow), while orphan peaks remain red and green.

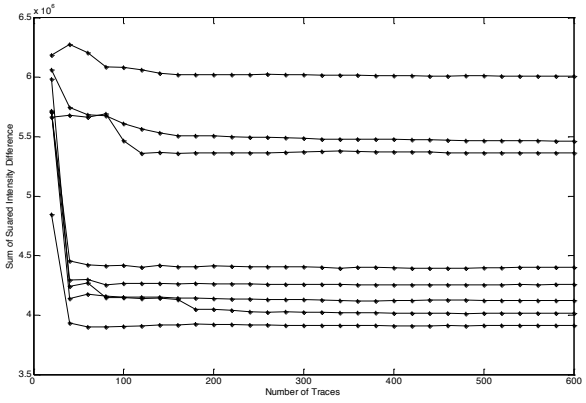

Figure S-3. Sum of squared intensity differences using all points in single-stage LC-MS images between the reference chromatogram and a sample chromatogram aligned using different numbers of high quality LCODA-selected mass traces using the PTW-CODA algorithm. The minimum sum of squared intensity differences is reached at around 200 selected mass traces in 9 pairs of chromatograms randomly selected from the urine dataset.

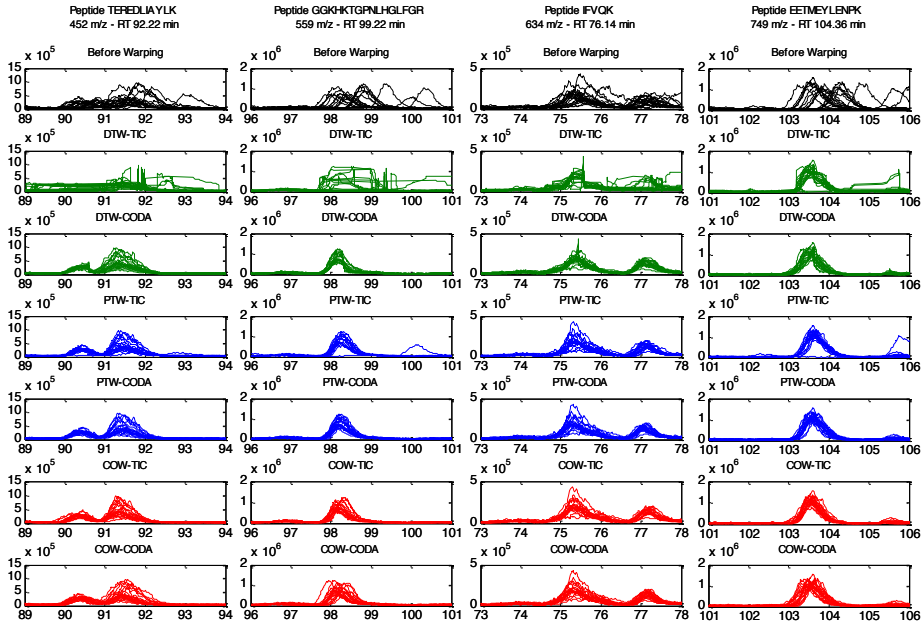

Figure S-4. Extracted ion chromatograms of spiked internal standard peptides, which are present in all trypsin-digested serum samples (cervical cancer data set, 20 chromatograms). Each column corresponds to one standard peptide and each row corresponds to a different time alignment algorithm with the following order from top to bottom: original dataset, DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA. The original dataset is in black, the DTW based algorithms are in green, PTW based algorithm in blue and COW based algorithms are in red. All time alignments were obtained with respect to the best reference chromatogram. All time alignment algorithms, except for DTW-TIC, are able to correct retention time shifts observed in the original dataset. Major misalignments and peak distortions are observed for all spiked standard peptides using DTW-TIC.

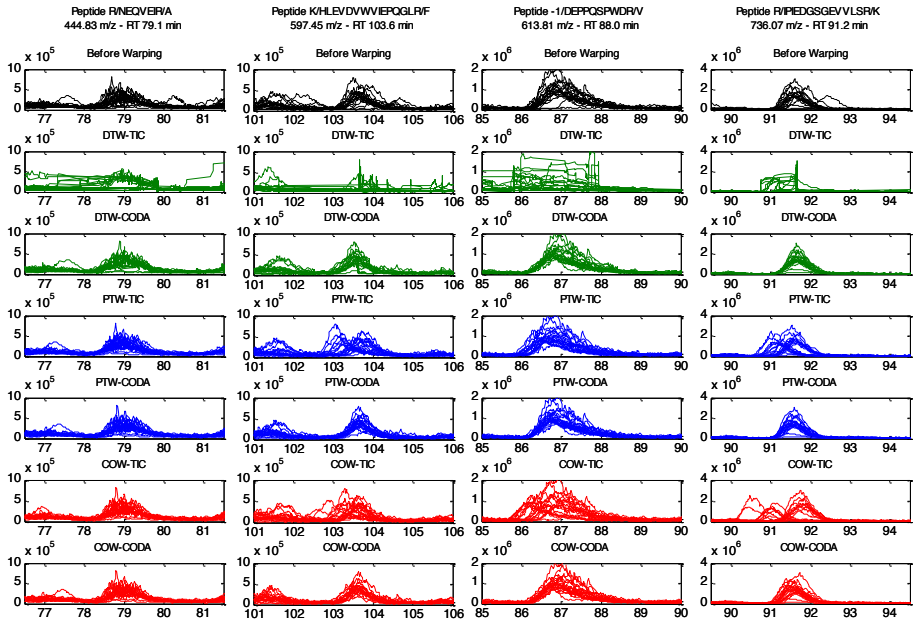

Figure S-5. Extracted ion chromatograms of spiked internal standard peptides, which are present in all serum samples of factorial design data set (19 chromatograms). Each column corresponds to one standard peptide and each row corresponds to a different time alignment algorithm with the following order from top to bottom: original dataset, DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA. The EIC's of corresponding mass traces in the original dataset are in black, the DTW based algorithms are in green, PTW based algorithm in blue and COW based algorithms are in red. All time alignments were obtained with respect to the best reference chromatogram. Major peak distortions and misalignments are observed for all peptides using the alignment with DTW-TIC. Two peptides are misaligned using COW-TIC and PTW-TIC. Improvements in the time alignment are clearly visible using algorithms combined with CODA- or LCODA-selected mass traces. The most striking differences in the time alignment performance can be observed for DTW-CODA where peak distortions are significantly reduced resulting in tightly aligned peaks.

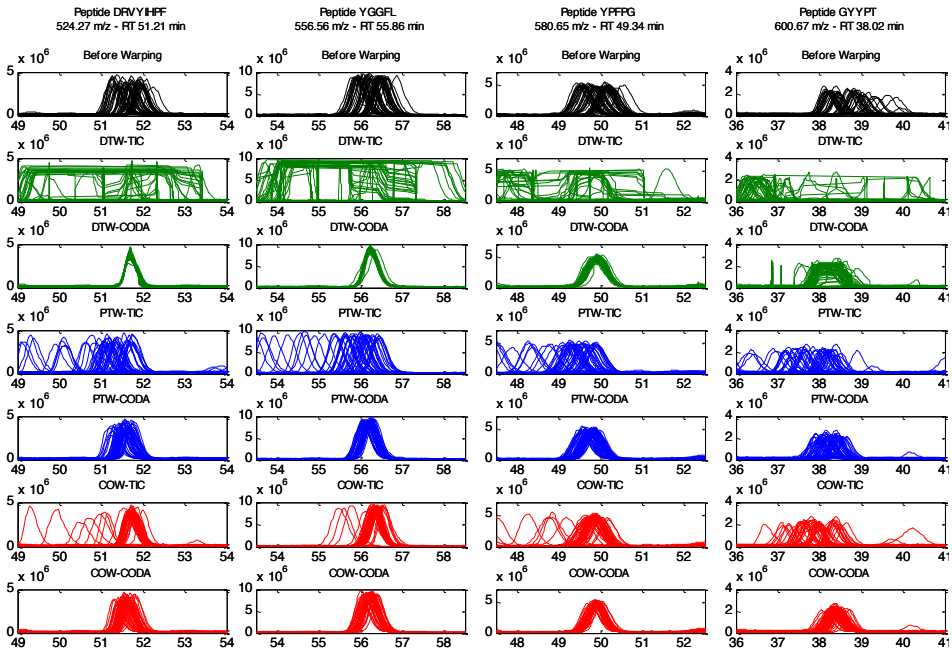

Figure S-6. Extracted ion chromatograms of spiked internal standard peptides, which are present in all acid-precipitated urine samples (urine data set, 50 chromatograms). Each column corresponds to one standard peptide and each row corresponds to the different time alignment algorithm with the following order from top to bottom: original dataset, DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA. The EIC's of corresponding mass traces in the original dataset are in black, the DTW based algorithms are in green, PTW based algorithm are in blue and COW based algorithms are in red. All time alignments were obtained with respect to the best reference chromatogram. Major misalignments are observed for all peptides using the alignment with COW-TIC, PTW-TIC, and DTW-TIC. Significant improvements in time alignment of standard added peptides were observed for all algorithms combined with CODA- or LCODA-selected mass traces. However some misalignments of the peptide GYPPT (the forth column) can still be observed even after alignment with PTW-CODA and DTW-CODA. Slight peak distortion can be observed for this peptide after alignment with DTW-CODA.

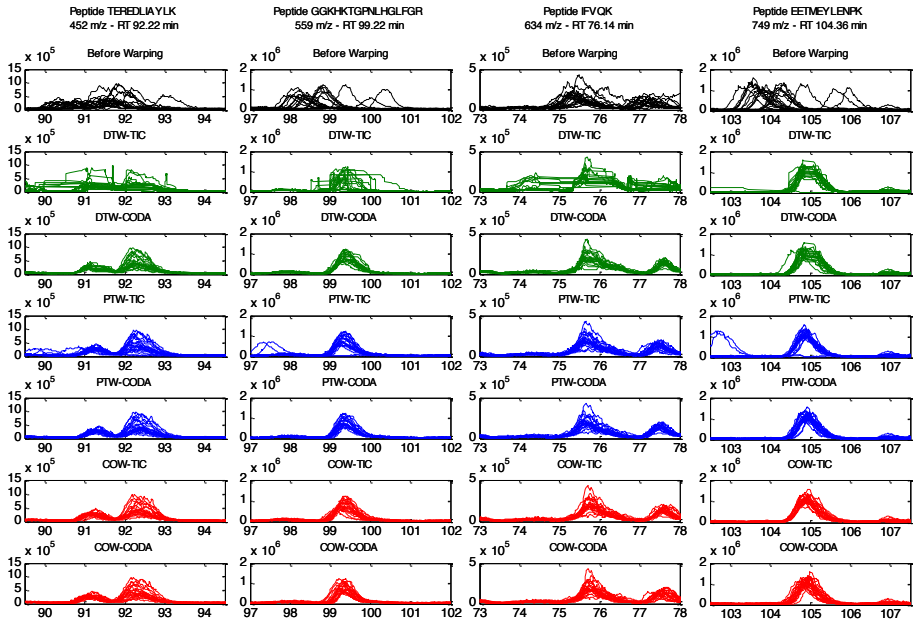

Figure S-7. Extracted ion chromatograms of spiked internal standard peptides, which are present in all trypsin-digested serum samples (cervical cancer data set, 20 chromatograms). Each column corresponds to one standard peptide and each row corresponds to a different time alignment algorithm with the following order from top to bottom: original dataset, DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA. The EIC's of corresponding mass traces in the original dataset are in black, the DTW based algorithms are in green, PTW based algorithm are in blue and COW based algorithms are in red. All time alignments were obtained with respect to the worst reference chromatogram. Major misalignments are observed for all peptides using DTW-TIC and misalignment for two peptides is visible in EIC obtained using PTW-TIC. COW-TIC, COW-CODA, PTW-CODA and DTW-CODA show improved time alignment for all peptides compared to the original data.

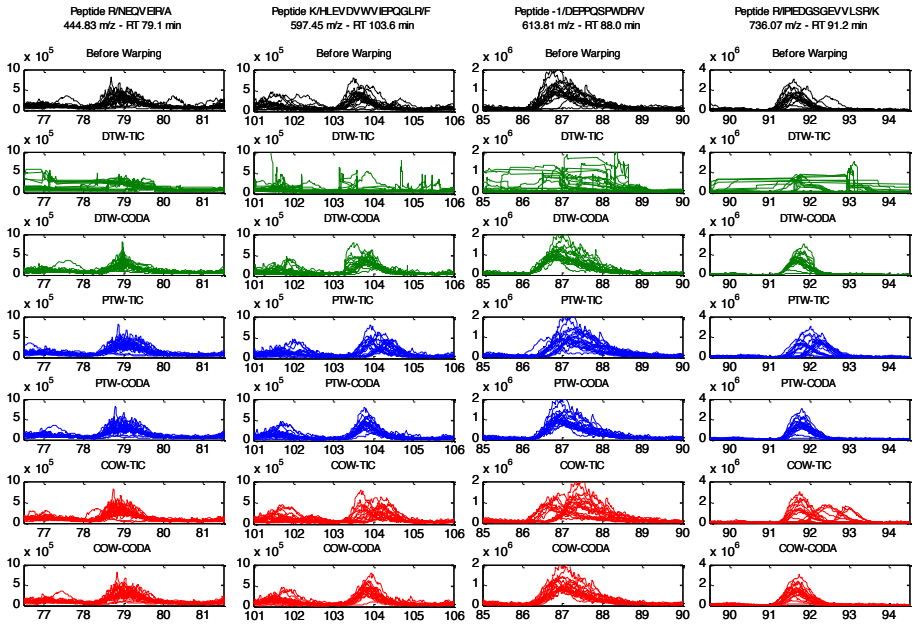

Figure S-8. Extracted ion chromatograms of spiked internal standard peptides, which are present in all serum samples of factorial design data set (19 chromatograms). Each column corresponds to one standard peptide and each row corresponds to a different time alignment algorithm with the following order from top to bottom: original dataset, DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA. The EIC's of corresponding mass traces in the original dataset are in black, the DTW based algorithms are in green, PTW based algorithm are in blue and COW based algorithms are in red. All time alignments were obtained with respect to the worst reference chromatogram. Major misalignments are observed for all peptides using DTW-TIC, which also results in considerable peak distortions. Misalignment for three peptides is visible in EIC obtained using PTW-TIC and COW-TIC. COW-CODA, PTW-CODA and DTW-CODA show improved time alignment for all peptides compared to the original data, although DTW-CODA still shows minor peak distortion (right column, peptide R/IPIEDGSGEVVLSR/K).

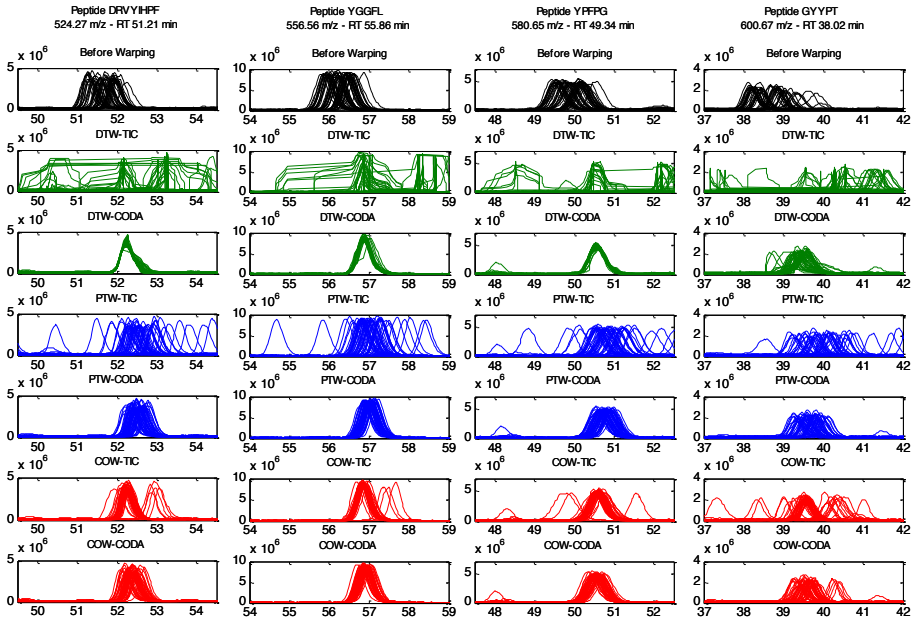

Figure S-9. Extracted ion chromatograms of spiked standard peptides, which are present in all acid-precipitated urine samples (urine data set, 50 chromatograms). Each column corresponds to one standard peptide and each row corresponds to a different time alignment algorithm with the following order from top to bottom: original dataset, DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA. The EIC's of corresponding mass traces in the original dataset are in black, the DTW based algorithms are in green, PTW based algorithm are in blue and COW based algorithms are in red. All time alignments were obtained with respect to the worst reference chromatogram. Major misalignments are observed for all peptides using DTW-TIC with considerable peak distortions. Misalignment for all peptides is visible in EIC obtained using PTW-TIC and COW-TIC. COW-CODA, PTW-CODA and DTW-CODA show improved time alignment for all peptides compared to the original data except for the GYYPT (right column). This peptide is one of the first eluting peaks, and has a low number of neighbor peaks in other mass traces with close retention time that could contribute to drive the local time alignment.

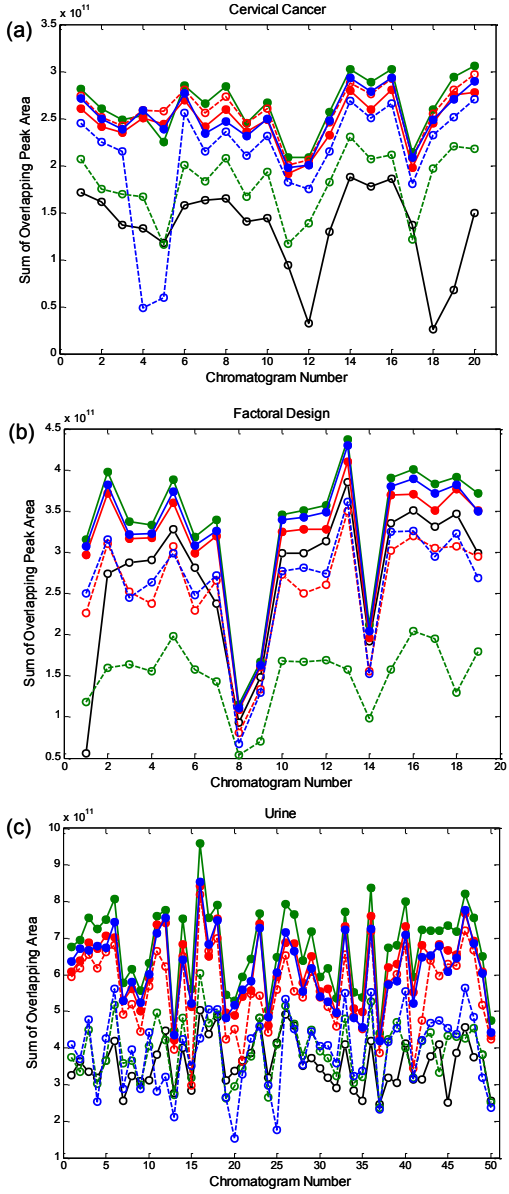

Figure S-10. Sum of overlapping peak areas of all chromatogram pairs using the worst reference chromatogram after applying M-N rules as peak filter to the cervical cancer (a), factorial design (b), and urine (c) data sets. The original chromatograms before alignment (black) are compared to the chromatograms obtained after alignment with COW (red), PTW (blue) and DTW (green) using TICs (dashed lines, empty circles) or CODA-/LCODA-selected mass traces (full lines, full circles). The chromatogram names corresponding to the chromatogram indices in the figures are reported in Supporting information (Table S2, S3, S4).

TABLES

Data set	Best Reference Chromatogram	Worst Reference Chromatogram
Cervical Cancer	17060511	2006052
Factorial Design	16090537	16090542
Acid-Precipitated Urine	05082628	05090150

Table S-1. File name of the chromatograms selected as the best and the worst reference from the three data sets.

Chromatogram Index	Chromatogram Name	Chromatogram Index	Chromatogram Name
1	17060510	11	1706052
2	17060511	12	1706053
3	17060512	13	1706054
4	17060513	14	1706055
5	17060514	15	1706056
6	17060515	16	1706057
7	17060516	17	1706059
8	17060517	18	2006051
9	17060518	19	2006052
10	17060521	20	2006053

Table S-2. File name of the chromatograms in cervical cancer data set.

Chromatogram Index	Chromatogram Name	Chromatogram Index	Chromatogram Name
1	16090524	11	16090534
2	16090525	12	16090535
3	16090526	13	16090536
4	16090527	14	16090537
5	16090528	15	16090538
6	16090529	16	16090539
7	16090530	17	16090540
8	16090531	18	16090541
9	16090532	19	16090542
10	16090533		

Table S-3. File name of the chromatograms in factorial design data set.

Chromatogram Index	Chromatogram Name	Chromatogram Index	Chromatogram Name
1	5082602	26	5082629
2	5082603	27	5082630
3	5082604	28	5090131
4	5082605	29	5090132
5	5082606	30	5090133
6	5082607	31	5090135
7	5082608	32	5090136
8	5082609	33	5090137
9	5082610	34	5090138
10	5082611	35	5090139
11	5082613	36	5090140
12	5082614	37	5090141
13	5082615	38	5090142
14	5082616	39	5090143
15	5082617	40	5090144
16	5082618	41	5090146
17	5082619	42	5090147
18	5082620	43	5090148
19	5082621	44	5090149
20	5082622	45	5090150
21	5082624	46	5090151
22	5082625	47	5090152
23	5082626	48	5090153
24	5082627	49	5090154
25	5082628	50	5090155

Table S-4. File name of the chromatograms in urine data set.

Appendix D

Supporting Information
Chapter 4

1 SAMPLE PREPARATION

LC-MS data acquisition

Sample preparation was performed as previously described [1]. The amount of urine injected into the LC-MS system was normalized to 50 nmol of creatinine. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

Preparation of spiked urine samples

600 µL carbonic anhydrase (CA) solution of 22 mg/mL dissolved in 50 mmol/L NH_4HCO_3 buffer at pH 7.8 were divided into 6 equal aliquots. Ten µL of 100 mM DTT were added to each aliquot and the solution was incubated at 50°C for 30 min followed by addition of 40 µL of 137.5 mM iodoacetamide and incubation at room temperature for another 60 min. Reduced and alkylated CA was digested by adding 40 µL of 0.5 µg/µL trypsin and subsequent incubation at 37°C over night. The reaction was stopped by addition of 10 µL pure formic acid (FA). The excess of DTT and iodoacetamide was removed by solid-phase extraction using a 100 mg Strata C-18 SPE column with the following protocol: the column was conditioned with 2 mL methanol, followed by one washing step with 2 mL water. Each aliquot of digested CA was loaded on the SPE column and the column was subsequently washed with 2 mL of 5% aq. methanol. Peptides were eluted with 1 mL of 80% aq. methanol. The eluate was dried in a vacuum centrifuge and re-dissolved in 200 µL 30% acetonitrile (ACN) and 1% FA. Finally 500 µL of digested CA were mixed with 200 µL of a stock solution of the synthetic peptides resulting in a standard mixture stock solution with a calculated digested CA concentration of 240 µM and the following concentrations (in µM) for the 7 synthetic peptides: VYV, 83; YGGFL, 57; DRVYIHPF, 29; YPFPGPI, 46; YPFPG, 60; GYYPT, 54; and YGGWL, 57.

Reversed-Phase LC-MS

All LC□MS analyses were performed on an 1100 series capillary HPLC system equipped with a cooled autosampler (4°C) and an SL ion trap mass spectrometer (Agilent Technologies, Santa Clara, CA, United States). Samples were desalted on an Atlantis dC18 precolumn (Waters Corporation, Milford, MA, USA, 2.1 × 20 mm, 3 µm particles, 10 nm pores) using 0.1% FA in 5% ACN at a flow rate of 50 µL/min for 16 min. Compounds were back-flushed from the precolumn onto a temperature-controlled (25°C) Atlantis dC18 analytical column (1.0 × 150 mm, 3 µm particles, 30 nm pores) and separated over 90 min at a flow rate of 50 µL/min during which the percentage of solvent B (0.1% FA in ACN) in solvent A (0.1% FA in ultrapure H_2O) was increased from 5.0 to 43.6% (eluent gradient of 0.43%/min). Settings of the electrospray ionization interface and the mass spectrometer were as follows: nebulization gas, 40.0 psi N_2; drying gas, 6.0 L/min N_2; capillary temperature, 325°C; capillary voltage, 3250 V; skimmer voltage, 25 V; capillary exit voltage, 90 V; octapole 1 voltage, 8.5 V; octapole 2 voltage, 4.0 V; octapole RF voltage, 175 V; lens 1 voltage, −5 V; lens 2 voltage, −64.6 V;

trap drive, 67; scan speed, 5500 m/z s^{-1}; accumulation time 50 ms (or 30 000 ions); scan range, 100−1500 m/z; a Gaussian smoothing filter (width 0.15 m/z) was applied for each mass spectrum; rolling average was disabled, resulting in a rate of approximately 70 mass spectra per minute. Spectra were saved in profile mode.

Following the gradient, both columns were washed with 85% B for 5 min and equilibrated with 5% B for 10 min prior to the next injection. Different volumes of the standard mixture (CA digest plus peptides) were injected on the pre-column prior to injection of the pooled urine sample to obtain the desired final concentrations. The injection system was cleaned with 70% ACN after each injection and filled with 0.1% FA in 5% ACN. Mass spectrometry settings were optimized for detection of singly- and doubly-charged ions of DRVYIHPF without provoking upfront fragmentation. Raw data converted to mzXML format are available at http://tinyurl.com/statisticsComparison. After the LC-MS analysis, the raw LC-MS profile data was exported in mzXML format using CompassExport v1.3.6.

1. Kemperman, R. F., Horvatovich, P. L., Hoekman, B., Reijmers, T. H., Muskiet, F. A., and Bischoff, R. (2007) Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: method development, evaluation, and application to proteinuria. *J Proteome Res* 6, 194-206.

2 FIGURES AND TABLES

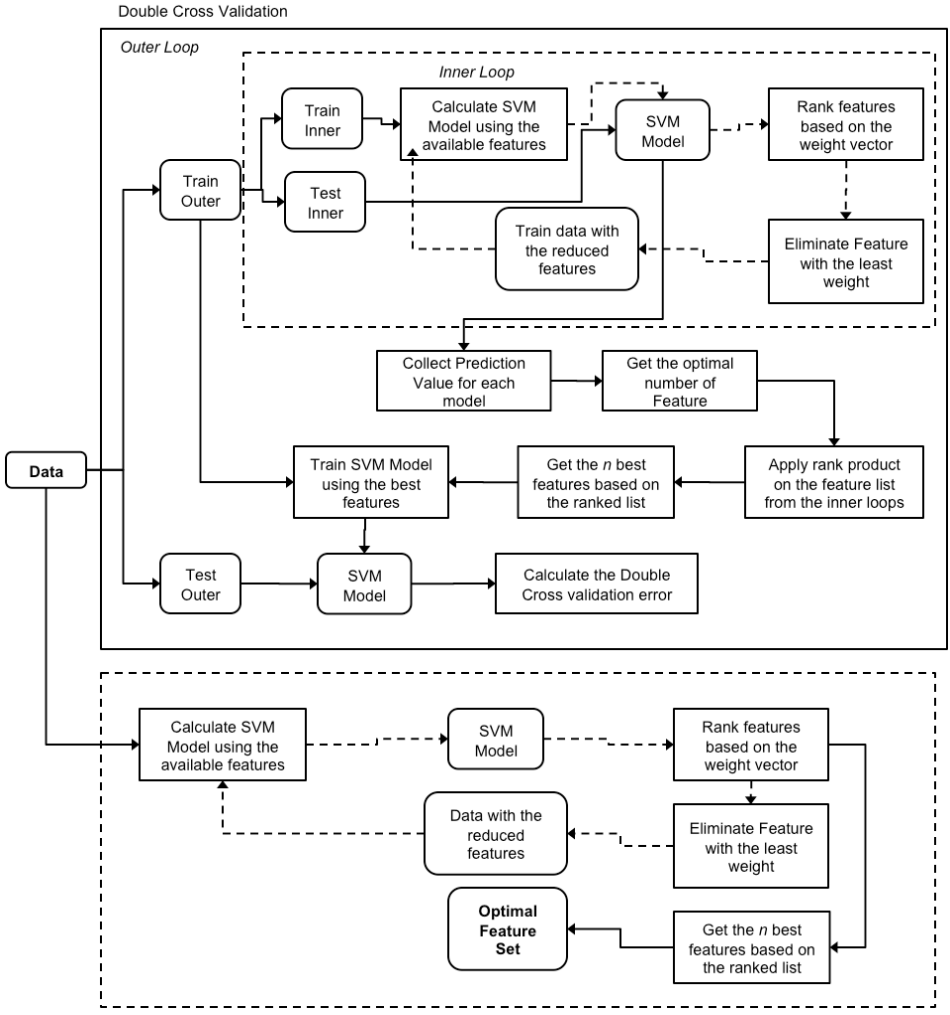

Figure S-1. Double cross validation scheme for a Support Vector Machine combined with Recursive Feature Elimination (SVM-FRE). The optimal number of features is obtained in the inner loops. The optimum model is then tested against the test data in the outer loop to obtain the overall classification error.

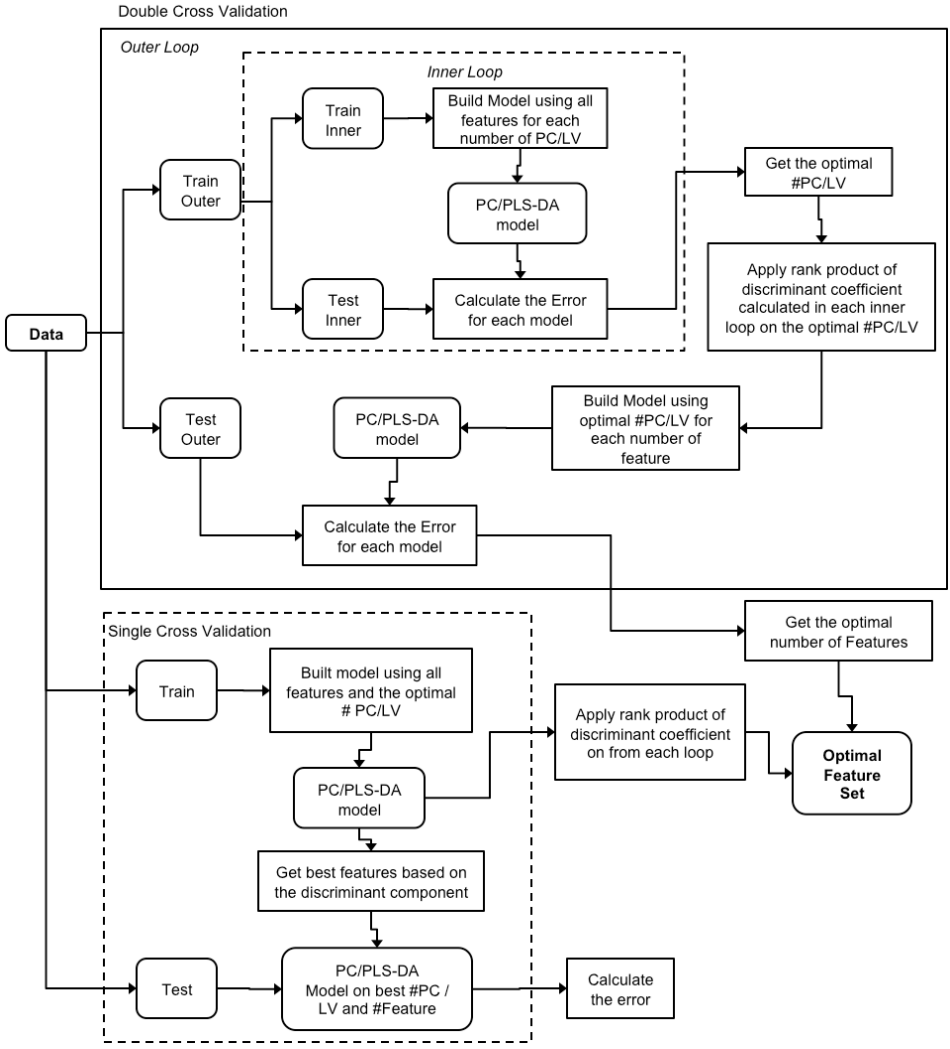

Figure S-2. Double cross validation scheme for PCDA and PLSDA. The optimal number of PC/PLS components is obtained in the inner loops. In the outer loop, the optimal number of features is determined by calculating the classification error for each ranked feature set. Once the optimal number of PC/PLS components and the optimal number of features has been obtained, a single cross validation scheme is performed to determine the optimal feature set. The optimum model is then tested against the test data in the outer loop to obtain the classification error.

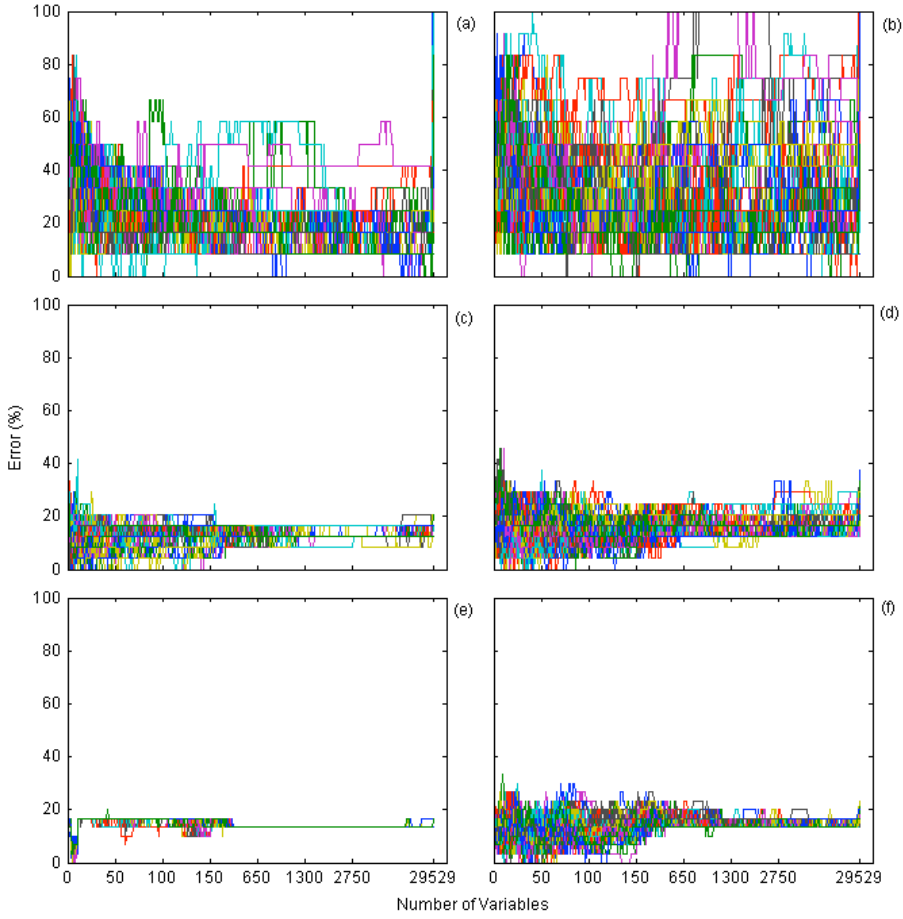

Figure S-3. Error plots of the PCDA model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets). The error plots show a decreasing variability with increasing sample size per class.

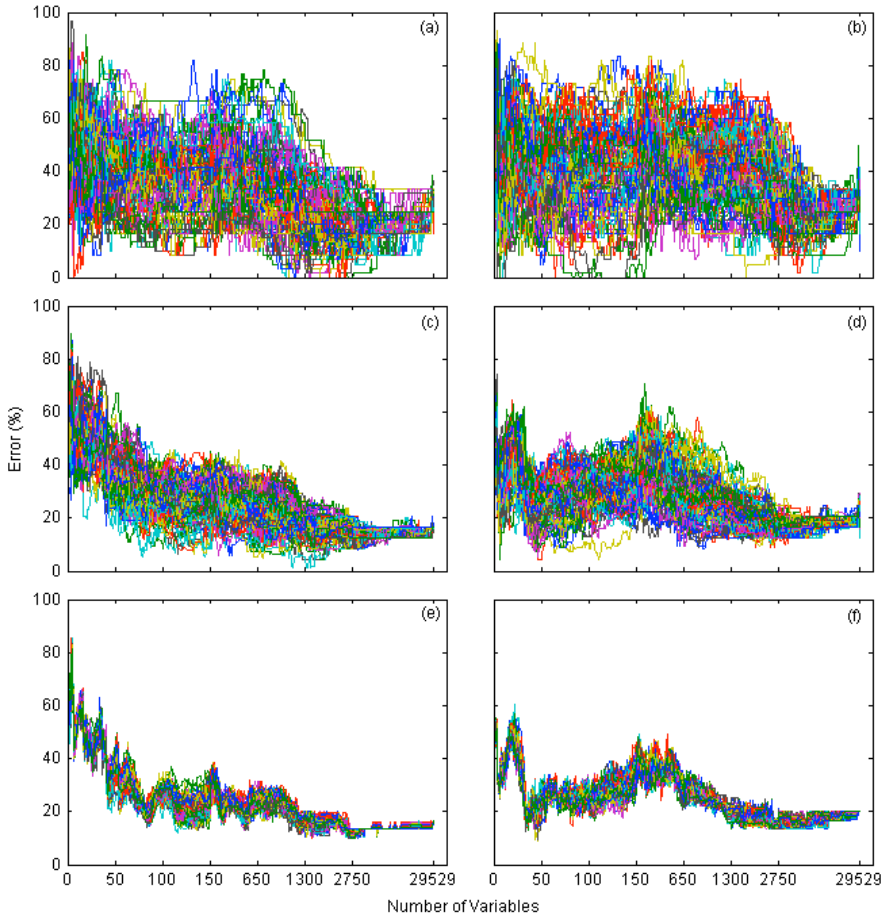

Figure S-4. Error Plot of the SVM-RFE model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets).

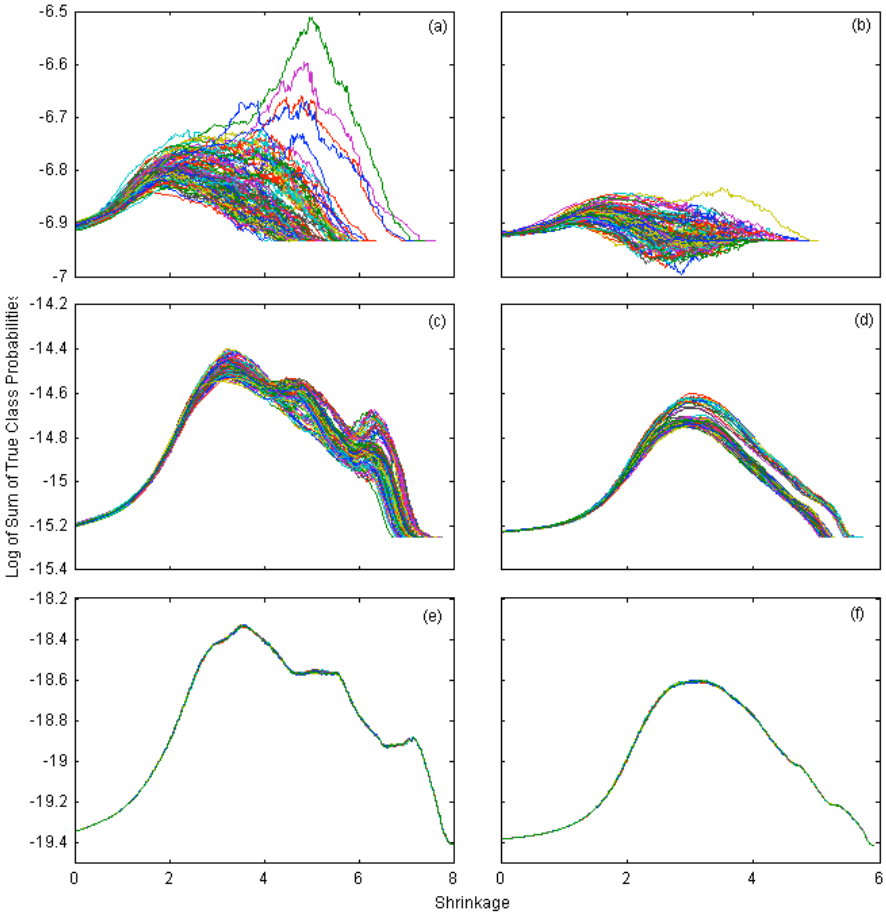

Figure S-5. Probability plot of the NSC model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets). The probability plots show a decreasing variability with increasing sample size per class.

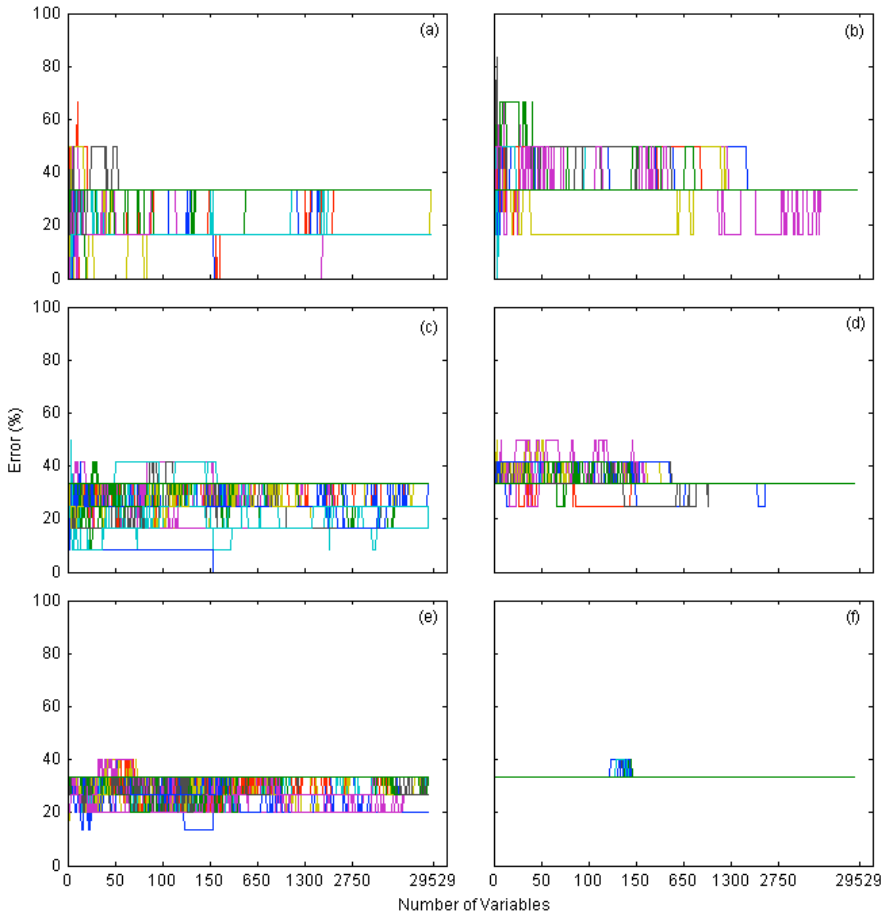

Figure S-6. Error plot of the PLSDA model based on 100 repetitions of the double cross validation scheme on data set 1a (a), data set 1b (c), data set 1c (e), data set 2a (b), data set 2b (d), and data set 2c (f) (see Table 1 for details about the data sets).

Methods: *mw*-test				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	0	0	0	0
1b	96	13	366	24
1c	108	108	274	274
2a	0	0	0	0
2b	22	2	34	3
2c	41	41	52	52
Methods: *t*-test				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	1	0	22	0
1b	101	18	348	23
1c	90	90	168	168
2a	1	0	17	0
2b	7	0	17	0
2c	39	39	46	46
Methods: NSC				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	141	1	3352	1
1b	87	47	143	53
1c	59	55	69	65
2a	137	6	10262	6
2b	49	25	82	25
2c	42	36	47	38

Methods: PCDA				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	51	0	28748	0
1b	134	0	857	0
1c	12	1	14	1
2a	151	0	29043	0
2b	89	0	394	0
2c	72	0	142	0
Methods: PLSDA				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	22	2	49	2
1b	77	2	425	2
1c	60	2	221	2
2a	46	2	1298	2
2b	51	2	2044	2
2c	7	2	12	2
Methods: SVM				
Data Set	*Unique TP*	*Common TP*	*Unique P*	*Common P*
1a	89	0	8428	2
1b	75	0	6356	70
1c	32	1	3236	84
2a	116	0	10756	4
2b	53	0	4331	7
2c	37	0	2976	33

Table S-1. Overview of the performance of different methods based on the ratio between unique true positives (Unique TP; selected at least once) and common true positives (Common TP; selected each time). The stability of the delivered feature set can be seen by comparing the number of unique features to the number of common features selected across each of the 100 repetitions (except for the *mww*-test and the *t*-test on data sets 1c and 2c, where repetitions were not possible, since all samples were used). Unique True Positive (Unique-TP) is a spiked-peptide-related feature that is selected at least once in 100 repetitions. Common True Positive (Common TP) a spiked-peptide-related feature that is always selected in each repetition. Unique Positive (Unique-P) is any feature that is included in a selected feature set at least once in 100 repetitions. Common Positive (Common P) is any feature that is always selected in each repetition.

List of Publications

Published

Christin C, Bischoff R, Horvatovich P. *Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery.* Talanta. 2011 Jan 30;83(4):1209-24.

Christin C, Smilde AK, Hoefsloot HC, Suits F, Bischoff R, Horvatovich PL. *Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms.* Anal Chem. 2008 Sep 15;80(18):7012-21.

Christin C, Hoefsloot HC, Smilde AK, Suits F, Bischoff R, Horvatovich PL. *Time alignment algorithms based on selected mass traces for complex LC-MS data*. Journal Proteome Res. 2010 Mar 5;9(3):1483-95.

Submitted

Christin C, Hoefsloot HC, Smilde AK, B. Hoekman, Bischoff R, Horvatovich PL. *A Critical Assessment of statistical methods for biomarker discovery in clinical proteomics.* Submitted.

Not related to this thesis

Rosenling T, Slim CL, Christin C, Coulier L, Shi S, Stoop MP, Bosman J, Suits F, Horvatovich PL, Stockhofe-Zurwieden N, Vreeken R, Hankemeier T, van Gool AJ, Luider TM, Bischoff R. *The effect of preanalytical factors on stability of the proteome and selected metabolites in cerebrospinal fluid (CSF).* Journal Proteome Res. 2009 Dec;8(12):5511-22.

Acknowledgement

I would not have accomplished this work without the help and support from so many people in various ways. I thank my promotors Rainer, Age and Ate for giving me the chance to start this project in the first place.

Thank you Rainer, for your supervision and your guidance during these years. You were not only a thesis promotor, your did not only care about scientific work but also in many aspects outside Ph.D. thesis. You were always there to help even when problems were not at all related to science.

I thank you Peter, for your 'day to day' supervision. Even though we did not always agree with each other, this thesis would never have come to an end without your supervision and for sure I learned a lot from you. I wish you the best for pursuing your professorship.

I thank Age and Huub, for welcoming me each time I needed help on my project. There was always a good atmosphere during the thesis meeting even when things got tough. To Huub and Suzanne, thanks a lot for revising the Dutch summary and even more for all the advices in the statistical part of my thesis.

I would like to thank the reading committee, Prof. dr. Kohlbacher, Prof. dr. Hankemeier, and Prof. dr. van den Heuvel, without whom I would not have been able to finalize the thesis. Thank you for your giving your feedback about my thesis within such a short time.

For the administration part, I would also like to thank the graduate school GUIDE: Riekje, Maaike, and Mathilda, and everybody in the International Student Desk, you always helped me with the paper work for my stay in Netherlands. And to our secretary Jolanda: you had helped me even before I arrived to Groningen, many thanks for that.

And I would like to thank all the colleagues in the Analytical Biochemistry group: Laurent, Robert, Theo, Jan Bosman, Mihaela, Ramses, Natalia, Julien, Eslam, Laurette, Lorenza, Therese, Vikram, Jos, Krisztina, Jan Willem, as well as all colleagues in the Pharmaceutical Analysis group. It is nice to know you all and thanks for the good moments. Especially to my office mates, Berendi and Ishtiaq, thanks for the good times in the office. ☺

Dear Diane, we only met a few times during my first year in Groningen, but you were willing to help me during my most difficult times. Your attention and care meant a lot for me. Thank you for everything, I will never forget that.

Dear Tejas, you have been here and there, supporting me through all the hard and good times during these four years, whether it is about thesis or 'whatsoever'. Without saying, you know how grateful I am to have known you. With regret, it is hard to forget that silly self-composed "poem" you used to recite whenever I needed some consolation, with a bit of luck, I do hope I can get over those words. I suggest you put it as one of your thesis propositions!

Thank you for my "Chalmers" friends: Darima, Bruno, Yoana, Janeli and err... Tejas. You have become my second family since I arrived to Sweden. Besides, I have received a lot of scientific input for my thesis during our chats, discussions, since we all do a Ph.D. on similar topics (why the heck did we do this to ourselves...?). We do not see each other too often anymore as we continue our lives in different places and I always look forward to travelling together again, although the scheduling is sometimes a hell by itself. And thanks a lot for Bruno, who helped a lot during these years. I wish you the best for your defense and the life after thesis!

To my "Stichers": Agni, Chinny, Ninta, Maya, Ari, Lerry, Ve, and Irene "Talawaz", we did not see each other for a long time, but still and all, it is always refreshing to hear your news, from time to time. Especially those rib-tickling skype conferences, they help to cure the attack of homesickness.

To Freddy, thanks a lot for everything for these past 12 years, who would have guessed that we would both finally end up in Europe pursuing a Ph.D. I wish you good luck with your thesis.

Thanks to my new colleagues in the Bioinformatics Laboratory for their kind support as I was learning my new job while at the same time finalizing my thesis. Thank you Antoine for you understanding when I had to urgently finish this book.

I thank all my friends from whom I received uncountable support during completing my thesis: Rosa "tinyocha", Pipit, Patricia, Retno, Durba, Daniela, Sadia, Livi, Laurent, Mylene, Mihaela, Julien, Bahram, Jan, Ramona, Sun-ho, Alinde, Habib Kazemi, Henk Meijer, Nico van der Sman, Abdou, Maria Lopez, Alicja, Therese and everybody, too many to be named. Thank you Livi and Mihaela for helping me whenever I needed help on English. And of course to my twin girls: Lory and Laurette, thanks for the lovely moments, I hope we will have a lot of those in the coming future! And next time we will manage not to burn the chocolate. ☺

Thanks to my Indonesian family in Groningen: Wisnu, Yuli, Joshua, Tita, Shanti, and Mira. It is hard to imagine passing these years without your friendship. It helped me in so many ways.

To my paranymphs: Tita and Tejas, this defense would have never been completed without your hard work. Thanks for accepting to be my paranymphs even though you are also busy with a lot of deadlines in completing your own theses. I hope I can be of some help for your thesis defenses later on as good as you are for mine.

A la famille Abello: Véronique, Alain, Huguette, Pierrette, Tatie, Annabelle, Monique et Jean-Pierre. Passer du temps avec vous, pendant les vacances de Noël, m'a toujours aidé à recharger les batteries et à régénérer mon esprit durant les années de thèse. Et le plus important est de me rappeler que j'ai une famille joyeuse en Europe, dont je fais partie. Merci, je vous aime tous.

Aku tidak akan pernah bisa meraih semua ini tanpa papa, mama dan dede. Tidak ada kata-kata yang bisa mengungkapkan betapa berharganya kalian dalam hidupku. Terima kasih atas kasih sayang, dorongan, doa-doa dan semuanya, sejak dari dimulainya perjalanan hidup ini, setiap hari yang kulalui, hingga saat akhir nanti. Kalian adalah alasan utama bagiku untuk selalu berusaha untuk melakukan yang terbaik.

Last but not least, to my Nico: between the peaks and valleys of those LC-MS chromatograms, I found you. And here we are, having gone through your thesis defense and mine, the future days can only get better. Thank you for your priceless support, for making the unbelievable things happen, and for your understanding during the hard and stressful moments in this thesis time. I can't wait for the next adventures in our lives and for the next best things to come. Ordinary but true, I love you!